인지과학, 제22권 제1호
Korean Journal of Cognitive Science
2011, Vol. 22, No. 1, 1～18.

휴대전화에서 단문 메시지로부터 일정 자동 등록

김 재 훈† 김 형 철

한국해양대학교 컴퓨터공학과

휴대전화가 대중에게 널리 보급됨에 따라, 단문 서비스(SMS)가 새로운 의사소통 수단으로 등장하고 있다. 단문 서비스는 가격이 저렴할 뿐 아니라 단문 메시지를 쉽게 저장할 수 있어 약속이나 일정 알림용으로 널리 사용된다. 본 논문은 단문 메시지로부터 일정 정보(시간과 장소)를 추출하여 자동으로 일정 관리 시스템에 등록하는 시스템을 개발한다. 단문 메시지는 짧고 간결하지만 비속어나 준말 등이 많이 사용된다. 이것이 일정 정보를 추출하는데 더욱 어렵게 한다. 또한 휴대전화에는 계산 능력과 저장 공간이 충분하지 못하므로 형태소 분석과 같은 일반적인 자연언어 처리 모듈을 그대로 사용하는 것은 다소 무리가 있다. 이 문제를 해결하기 위해서 본 논문에서는 형태소 분석과 같이 복잡한 언어 처리 모듈을 사용하지 않고 기계학습 기반으로 일정 정보를 추출하고 추출된 정보를 휴대전화의 일정 관리 시스템에 등록한다. 본 논문에서 제안된 자동 일정 등록 시스템은 삼성전자 옴니아 휴대전화에 탑재되어 정상적으로 잘 동작함을 확인할 수 있었다.

주제어 : 단문 메시지, 일정 정보 추출, 휴대전화 일정 관리

---

† 교신저자: 김재훈, 한국해양대학교 컴퓨터공학과, 연구분야: 자연언어처리
E-mail: jhoon@hhu.ac.kr

서 론

휴대전화가 대중에게 널리 보급되면서 단문 메시지가 의사 전달의 중요한 수단으로 등장하게 되었다. 단문 메시지는 통화료가 저렴하고 통화 내용을 바로 저장할 수 있다는 큰 장점을 가지고 있어 약속이나 일정을 정하는 등에 매우 자주 활용되고 있다. 개인에 따라 조금씩 차이는 있지만 대부분의 사람들은 하루에도 수십 건의 단문 메시지를 주고받으며 그 메시지의 약 18%는 약속이나 일정에 관련된 것이다[1]. 그러나 현재 대부분의 휴대전화는 단문 메시지 시스템과 일정 관리 시스템은 완전히 독립적인 시스템으로 동작되므로 약속이나 일정에 관련된 단문 메시지가 도착하더라도 이를 바로 일정 관리 시스템에 등록할 수 없다. 본 논문에서는 이와 같은 번거로움을 해소하기 위해 단문 메시지가 도착하면 약속이나 일정에 관련된 것이지를 판단하여 자동으로 일정 관리 시스템에 등록하는 시스템을 개발한다. 본 논문에서는 이 시스템을 자동 일정 등록 시스템이라고 한다.

단문 메시지로부터 약속이나 일정 정보를 추출하기 위한 몇 가지 문제점을 살펴보자. **첫째,** 단문 메시지는 매우 비문법적이므로 일반적인 언어처리 모듈의 거의 사용할 수 없다. 일반적으로 단문 메시지는 매우 짧고 간결하지만 문법에 어긋나는 경우와 비속어, 신조어, 사투리 등을 사용하는 경우가 매우 자주 발생한다. 또한 80 바이트 내에 자신의 의사를 전달하려고 말을 줄이거나 띄어쓰기를 생략하는 경우가 매우 자주 발생한다. 이러한 문제 때문에 일반적인 언어처리 엔진을 거의 사용할 수 없다. 이 문제를 해결하기 위해 본 논문에서는 형태소나 단어 대신에 음절 2-그램(bigram)을 이용한다. 음절 2-그램은 정보검색 분야에서 많이 사용되었으며, 그 결과도 어느 정도 만족스러웠다[2-3]. **둘째,** 일반적인 정보추출에 비해서 학습자료 부족 문제(data sparseness problem)가 매우 심각하다. 일반적으로 정보추출을 위해서는 기계학습 방법이 많이 사용되고[4], 기계학습을 사용할 경우 많은 양의 학습자료를 필요로 한다. 그러나 단문 메시지는 개인 정보가 많이 포함되어 있어 쉽게 수집할 수 없기 때문에 학습자료 부족 문제가 더욱 심각하게 발생된다. 이 문제를 해결하기 위해 본 논문에서는 본 연구팀의 구성원들이 3개월 동안 사용한 6,788건의 단문 메시지를 수집하여 말뭉치를 구축하였다(3장 참조). **셋째,** 일반 PC에 비해 휴대용 단말기들은 계산 능력뿐 아니라 저장 용량도 매우 떨어진다. 이

를 해결하기 위해 본 논문에서는 최대한 자원을 적게 사용하며 적절한 성능을 발휘하는 기계학습 도구(CRFsuite)[5]를 사용하였으며, 또한 시간과 장소에 관련된 사전 정보 이외에 부가적인 언어 정보는 거의 사용하지 않았다. 이를 종합하여 제안된 자동 일정 등록 시스템은 삼성전자의 옴니아(Omnia) 휴대전화에 탑재되어 정상적으로 잘 동작함을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 정보추출과 CRF 모델에 대해서 간단히 소개하고 3장에서는 제안된 시스템을 자세히 설명한다. 4장에서 구현 및 성능 평가에 대해서 기술하고, 마지막으로 5장에서 결론을 맺고 앞으로의 연구과제에 대해서 기술한다.

관련 연구

정보추출

대규모의 데이터로부터 새로운 지식(이익이 되거나 관심이 있는 규칙이나 패턴)을 찾아내는 과정을 데이터 마이닝(data mining)이라고 하고, 대규모의 텍스트로부터 새로운 지식을 찾아내는 과정을 텍스트 마이닝(text mining)이라고 한다. 정보추출(information extraction)은 텍스트 마이닝의 한 분야로서 텍스트로부터 개체명(named entity)을 찾고, 찾아진 개체명들 사이의 관계를 추론하는 것이다[6]. 단문 메시지에서 약속 시간 및 장소를 추출하는 것도 정보추출의 한 예로 볼 수 있다. 최근 대부분의 정보추출 시스템은 기계학습을 기반으로 구현되었으며, 그 방법으로 크게 규칙 학습 기반 방법(rule learning based method)[7], 분류 모델 기반 방법(classification model based method)[8-9], 순서정보를 이용한 라벨 부착 방법(sequential labeling based method)[10-12]이 있다. 각각의 방법마다 특색이 있고 어떤 것이 더 효율적인 방법이라고 단언할 수는 없으나 정보추출에는 순서정보를 이용한 라벨 부착 방법이 널리 사용된다.

CRF(Conditional Random Fields)

CRF는 품사 부착과 같이 연속적인 자료에 라벨을 결정할 때, 매우 유용한 분별 확률 모델(discriminative probability model)이다(식 (1)). 즉 주어진 입력 벡터 $\mathbf{x}$에 대해

서 조건부 확률 $p(\mathbf{y}|\mathbf{x})$를 최대로 하는 라벨 $\mathbf{y}^*$를 선택하는 비방향성 그래프 모델(undirected graph model)이다[12].

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \tag{1}$$

여기서 $p(\mathbf{y}|\mathbf{x})$는 CRF의 종류에 따라 다양하게 정의될 수 있다. 본 논문의 경우에는 선형 연쇄 모델(linear chain model)을 적용하였으며 식 (2)과 같이 정의한다.

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right) \tag{2}$$

여기서 $f_k(y_t, y_{t-1}, \mathbf{x}_t)$는 자질 함수(feature function)이며, 자질 k에 따른 특성 함수(characteristic function)이다. 즉 주어진 입력 $y_{t-1}, y_t, \mathbf{x}$에 자질 k가 포함되어 있으면, 1을 반환하고 그렇지 않으면 0을 반환한다. λ_k는 매개변수이며 자질 k의 가중치가 된다. λ_k의 학습 방법은 일반적으로 기울기 하강 알고리즘(gradient descent algorithm)[8][13]과 준뉴턴 방법(quasi-Newton method)[14]를 주로 사용한다. $Z(\mathbf{x})$는 정규화 요소이다.

CRF 모델은 정보추출의 많은 분야에서 널리 사용되고 있으며, 특히 많은 자질과 다양한 자질 정보를 요구하는 응용분야에 적합한 모델이다[12]. 구체적인 응용분야를 살펴보면, 개체명 인식(named entity recognition)[15], 참조해소(coreference resolution)[16], 전문용어 인식(term recongnition)[17] 등 비교적 복잡한 문제를 해결하기 위해 두루 사용된다.

단문 메시지로부터 자동 일정 등록 시스템

자동 일정 등록 시스템의 핵심은 단문 메시지로부터 약속 시간과 장소를 추출하는 것이다. 이를 위해서 본 논문에서는 순서정보를 이용한 라벨 부착 방법인 CRF 모델을 이용한다. CRF 모델을 위한 학습 도구는 여러 종류가 공개되어 있다

[5][18]. 휴대전화의 경우에는 계산 능력과 저장 용량을 제한되므로 계산량과 저장 용량이 작은 CRFsuite를 사용하였다[19]. 약속 시간과 장소 정보를 추출하기 위해 먼저 자동 띄어쓰기를 수행하는데 이 시스템 또한 CRF 모델을 이용한다. 각 시스템에서 사용되는 자질집합(feature set) 등은 3.4절과 3.5절에서 자세히 기술한다. 일반적으로 기계학습 기반 시스템은 학습 시스템과 인식 시스템으로 구성된다[6]. 학습 시스템은 단문 메시지로부터 자질(feature)을 추출하고(3.4절과 3.5절 참조) 추출된 자질을 이용해서 학습 모델을 생성하며 휴대전화와는 독립적으로 수행된다. 인식 시스템은 입력된 단문 메시지로부터 자질을 추출하고 추출된 자질과 학습 모델을 이용해서 원하는 정보를 추출하며 휴대전화에서 수행되어야 한다. 또한 기계학습 기반 시스템은 많은 양의 학습 말뭉치가 필요하다. 본 논문에서는 본 연구팀의 구성원들이 3개월 동안 사용한 6,788건의 단문 메시지를 수집하여 말뭉치를 구축하였다(3.3절). 이하에서는 먼저 전체 시스템 구성에 대해서 간략히 기술하고 각 구성요소에 대해서 자세히 설명한다.

전체 시스템 구성

단문 메시지로부터 자동 일정 등록 시스템의 전체 흐름은 그림 1과 같다. 휴대

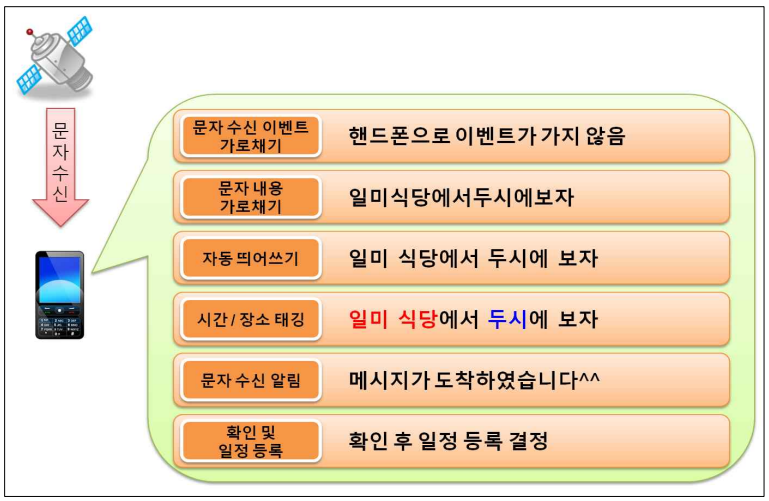

그림 1. 자동 일정 등록 시스템의 전체 구성도

전화에 단문 메시지가 수신되면 실제 휴대 전화의 단문 메시지 처리 시스템으로 전달되지 않도록 이벤트를 가로챈다(hooking). 이렇게 얻어진 메시지에 대해 자동 띄어쓰기를 수행하고, 그 결과를 이용해 약속 시간/장소 정보를 추출한다. 추출된 일정은 본 논문에서 구현된 GUI에 전달되어 사용자가 확인하고 자동으로 일정에 등록된다.

LGT 오즈 옴니아에서의 메시지 가로채기

윈도우즈 모바일(Windows Mobile)이 탑재된 스마트폰에서 메시지를 가로채려면 기본적으로 MS-SMS(Microsoft System Management Server)[20]를 사용하여야 한다. MS-SMS란 윈도우 모바일에서 제공해 주는 단문 메시지 서비스인데, 우리나라의 경우 보통 각 통신사에서 제작한 단문 메시지 서비스를 이용하도록 기본 옵션이 있다. 이러한 각 통신사의 송수신 서비스를 사용하지 않고 MS-SMS를 활성화하는 방법은 통신사마다 조금씩 다르며 **LGT** 오즈 옴니아의 경우는 다음과 같다(그림 2).

1. Iconsoft PhoneEX[21]를 설치한다.
2. Outlook을 활성화시킨 후 \Windows\kmonitor.exe를 실행하여 LGT Message Router을 중지시킨다.
3. WindowsMobile.PocketOutlook의 MessageInterception namespace에서 MessageInterceptorEventHandler를 이용하여 사용자에게 수신된 메시지를 가로챈다.
4. Notify와 NotifyAndDelete 옵션을 사용하여 수신된 메시지를 사용자에게 그대로 전달할지를 결정한다.

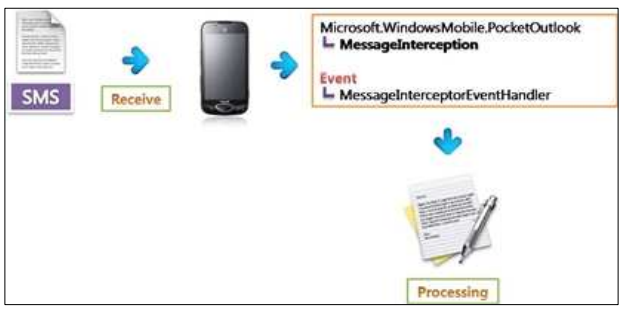

그림 2. 단문 메시지의 가로채기 과정

일정 추출을 위한 말뭉치 구축

기계학습을 사용할 경우 많은 양의 학습자료를 필요로 한다. 그러나 단문 메시지는 개인 정보가 많이 포함되어 공개 말뭉치가 없을 뿐 아니라 수집 작업도 쉽지 않다. 이 문제를 해결하기 위해 본 논문에서는 본 연구팀의 구성원들이 3개월 동안 사용한 6,788건의 단문 메시지를 수집하여 말뭉치를 구축하였으며 각 메시지는 수동으로 띄어쓰기를 수정하여 약속 시간과 장소 정보를 부착하였다(표 1).

표 1에서 보는 바와 같이 단문 메시지는 많은 경우 띄어쓰기를 제대로 하지 않으므로 본 논문에서는 수동으로 수정하여 말뭉치를 구축하였다. 띄어쓰기 정보는 음절 단위로 0이나 1을 부착하며 0은 띄어쓸 경우이고 1은 붙여쓸 경우이다. 시간과 장소 정보는 L과 T 태그를 부착하며 L 태그는 장소(Location)를 나타내고, T 태그는 시간(Time)을 나타낸다.

자동 띄어쓰기 시스템의 자질집합

단문 메시지는 80 바이트 내에 자신의 의사를 전달하려고 말을 줄이거나 띄어쓰기를 생략하는 경우가 매우 자주 발생한다. 따라서 본 논문에서는 자동 띄어쓰기 시스템을 통해서 띄어쓰기를 복원한다. 띄어쓰기는 글의 가독성을 높이고 문장의 뜻을 명확히 전달하기 위해 매우 중요하다. 실질적으로 자동 띄어쓰기 시스템을 사용할 경우가 뛰어쓰기를 수행하지 않았을 경우보다 더 좋은 성능을 보였다

표 1. 학습 말뭉치의 구성

구 분	단문 메시지 #1	단문 메시지 #2
원본	야 대영에서2시늦지마라	오늘부산역에서볼꺼얌?
띄어쓰기 수정	야 대영에서 2시 늦지마라	오늘 부산역에서 볼꺼얌?
띄어쓰기 정보 부착	야/1 대/0 영/0 에/0 서/1 2/0 시/1 늦/0 지/0 마/0 라/1	오/0 늘/1 부/0 산/0 역/0 에/0 서/1 볼/0 꺼/0 얌/0 ?/1
약속 장소/시간 정보 부착	야 <L>대영</L>에서 <T>2시</T> 늦지 마라	<T>오늘</T> <L>부산역</L>에서 볼꺼얌?

표 2. 자동 띄어쓰기를 위한 자질 집합

자질 이름	적용 범위				비 고			
음절	s_{i-2}	s_{i-1}	s_i	s_{i+1}	s_{i+2}	s_{i-1}/s_i	s_i/s_{i+1}	
예	$	$	**야**	대	영	$/야	야/대	부류 1
	$	야	**대**	영	에	야/대	대/영	부류 0
	야	대	**영**	에	서	대/영	영/에	부류 0
	대	영	**에**	서	2	영/에	에/서	부류 0
	영	에	**서**	2	시	에/서	서/2	부류 1
	에	서	**2**	시	늦	서/2	2/시	부류 0
	서	2	**시**	늦	지	2/시	시/늦	부류 1
	2	시	**늦**	지	마	시/늦	늦/지	부류 0
	시	늦	**지**	마	라	늦/지	지/마	부류 0
	늦	지	**마**	라	$	지/마	마/라	부류 0
	지	마	**라**	$	$	마/라	라/$	부류 1

(표 11 참조). 자동 띄어쓰기 시스템도 앞에서 언급했듯이 기계학습을 기반으로 구현되며, 표 2는 띄어쓰기에 위한 자질집합과 표 1의 메시지 “야 대영에서2시늦지마라”에 대한 자질집합의 예이고 ‘$’는 공백을 의미한다.

일정 추출 시스템의 자질집합

일정 추출을 위한 학습의 기본 단위는 음절 2-그램(syllable bigram)을 사용하며 본 논문에서는 편의상 토큰(token)이라고 한다. 띄어쓰기의 경우와는 다르게 하나 이상의 토큰이 결합되어야 최종 시스템의 결과(약속 시간/장소)를 출력할 수 있다. 예를 들면 ‘남포동’이라는 장소는 ‘남포’와 ‘포동’이라는 두 개의 토큰이 결합되어야 한다. 이처럼 하나 이상의 토큰이 결합되어 개체명이나 복합명사를 구성할 때 널리 사용되는 부호화 방법이 BIO 부호화(BIO encoding)[22]이다. BIO 부호화는 개체명이나 복합명사에 속하는 토큰을 구별하여 특별한 태그(tag)를 덧붙이는 방법이다. 즉 토큰이 개체명의 시작에 나타나면 태그 ‘B’를 덧붙이고 개체명의 내부에 나타나면 태그 ‘I’를 덧붙이는 방법이다. 예를 들면 표 3은 표 1의 학습 말뭉치로부터

표 3. 학습 말뭉치에 대한 BIO 부호화의 예

구 분	단문 메시지 #1			단문 메시지 #2		
약속 장소/시간 정보 부착	야 <L>대영</L>에서 <T>2시</T> 늦지 마라			<T>오늘</T> <L>부산역</L>에서 볼꺼얌?		
BIO 부호화	$야/O 야$/O $대/L-B 대영/L-I 영에/O	에서/O 서$/O $2/T-B 2시/T-I 시$/O	$늦/O 늦지/O 지_/O _마/O 마라/O 라/$/O	$오/T-B 오늘/T-I 늘$/O $부/L-B 부산/L-I	산역/L-I 역에/O 에서/O 서$/O $볼/O	볼꺼/O 꺼얌/O 얌?/O ?$/O

BIO 부호화한 예이다.

본 논문에서 일정은 약속 시간 및 장소를 의미하며 하나 이상의 토큰이 결합되어야 하나의 일정을 결정할 수 있다. 앞에서 언급했듯이 휴대전화의 여러 가지 제약으로 토큰(음절 2-그램)을 자질의 기본 단위로 사용하며 본 논문에서 사용하는 자질집합은 표 4과 같다.

표 4에서 첫 번째 자질은 형태소(혹은 단어)가 아닌 토큰 자체를 자질로 사용한 것이다. 현재 토큰(w_i)의 정답을 결정하기 위해서 앞에 나오는 2개의 토큰(w_{i-2}, w_{i-1})과 뒤에 나오는 2개의 토큰(w_{i+1}, w_{i+2})을 자질로 사용한다(전체 윈도우 크기: 5). 두 번째와 세 번째 자질은 미리 구축된 시간/장소 사전(표 5 참조)에 현재 토큰(w_i)의 출현 여부에 따라 '+'나 '–' 기호를 자질로 사용한다. 장소

표 4. 일정 추출을 자질집합

자질 이름	적용 범위	비 고
토 큰	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, w_{i-2}/w_{i-1}, w_{i-1}/w_i, w_i/w_{i+1}, w_{i+1}/w_{i+2}, $w_{i-1}/w_i/w_{i+1}$	토큰 그대로의 자질
시간 사전	w_i	출현 유무
장소 사전	w_i	출현 유무

사전의 경우, 말뭉치에 나타난 장소 정보와 네이버 지도[22]에서 나타난 위치 정보를 수집하여 만들었다. 시간 사전의 경우는 장소명과는 달리 신조어가 생성되기 힘들고, 나타날 수 있는 거의 대부분의 시간을 정리할 수 있기 때문에 본 연구팀이 직접 제작하였다. 각 사전은 말뭉치와 마찬가지로 토큰으로 표현되어 있으며 표 5에 시간 및 장소 사전의 예를 나타내었다.

표 5. 자질로 이용될 시간 및 장소 사전

구 분	표제어의 예
시간 사전	저녁, 정오, 제헌, 현절, 한시, 2시, 3시, ...
장소 사전	서면, 부산, 산역, 대신, 신동, 대전, 대구, ...

예를 들어, 메시지 “야 대영에서 2시 늦지마라”에서 “2시”를 w_i라고 할 때 자질 집합은 표 6와 같다.

표 6. 일정 추출을 위한 자질집합의 예

자질종류	토큰 자질						비고
	w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}	w_{i-2}/w_{i-1}	
자질값	서_	_2	**2시**	시_	_늦	서_/_2	부류 T-I
자질종류	토큰 자질		시간자질		장소자질		
	w_{i-1}/w_i	w_i/w_{i+1}	w_{i+1}/w_{i+2}	$w_{i-1}/w_i/w_{i+1}$	w_i	w_i	
자질값	_2/2시	2시/시_	시_/_늦	_2/2시/시_	+	-	

일정 등록

3.5절에서 약속 시간/장소가 인식되면 사용자에게 그 결과를 보여주고 휴대전화의 일정 관리 시스템에 일정을 등록하면 된다. 이 과정에서 약간의 문제가 발생된다. 장소는 고유명사이므로 그대로 등록하면 된다. 그러나 시간의 경우는 많은 경우가 상대적인 속성을 가지고 있다. 예를 들면 ‘오늘’, ‘내일’ 등은 상대적인 시간이다.

좀 더 구체적인 문제를 살펴보면 ‘2시’, ‘내일 2시’, ‘모래 2시’가 모두 같은 ‘2시’ 이지만 이들은 전혀 다른 시간이다. 따라서 이들을 구별할 수 있는 방법이 필요하다. 본 논문에서는 이러한 문제를 해결하기 위하여 일정 산출표를 이용한다(표 7).

표 7. 일정 등록을 위한 일정 산출표

약속 시간 개체명	절대 시간		상대 시간	
	날짜	시각	날짜	시각
오늘	-	-	0	0
내일	-	-	+1	0
모레	-	-	+2	0
세시	-	3	0	0
낼 모레	-	-	+2	0
닷새 뒤	-	-	+5	0
열달 후	-	-	+300	0
정오	-	12	0	0
두 시간 후	-	-	0	+2

표 7의 일정 산출표에서 시간은 절대 시간(‘오후 4시’, ‘오전 2시’, 정오 등)과 상대 시간(내일, 모레, 이틀 후 등)으로 구분되며 또 각 시간은 날짜(일)와 시각(시)으로 분리된다. 절대 시간은 인식된 약속 시간을 숫자로 변경하여 그대로 등록하면 된다. 상대 시간은 일정 산출표에서 얻은 값을 기준 날짜와 시각(오늘 현재 시각)에 더하면 절대 시간을 구할 수 있다. 예를 들어 10월 3일에 “내일 세시에 보자”라는 단문 메시지를 받으면 “내일 세시”라는 약속 시간이 인식되고 이는 상대 시간 ‘내일’에 의해 “10월 3일 + 1일 = 10월 4일”이 구해지고 절대 시간 ‘세시’에 의해서 “3시”가 구해져서 “10월 4일 3시”를 약속 시간으로 등록한다.

시스템 구현 및 성능 평가

시스템 구현 및 테스트

휴대전화의 경우에는 계산 능력과 저장 용량을 제한되므로 계산량과 저장 용량이 작은 CRF 학습 도구 CRFsuite를 사용하였다[19]. 본 논문에서 제안된 자동 일정 등록 시스템은 통신사로서 LGT를 선정하였고, 휴대 전화 기기는 삼성전자의 오즈 옴니아(OZ Omnia)를 선택하였다. 그림 3은 오즈 옴니아에서 실행과정을 화면 그대로 촬영된 것이며 동작에는 아무런 문제가 없었다.

(a) (b) (c) (d)

(a) 는 평상시 휴대 전화의 모습이고, (b) 는 단문 메시지가 수신된 모습이고,
(c) 약속 시간/장소가 인식된 모습이며, (d) 일정 등록이 완료된 모습이다.

그림 3. 오즈 옴니아에서 자동 일정 등록 시스템의 실행과정

학습 및 실험 말뭉치

앞에서 언급했듯이 본 논문에서 사용된 말뭉치는 본 연구팀에 의해서 직접 구축된 것이다. 표 8은 말뭉치의 구성을 나타내고 있다. 전체 말뭉치의 크기는 6,788건이며, 이 중에 일정(약속 시간과 장소)이 포함되지 않은 메지시가 3,564건이다.

표 8. 구축된 말뭉치의 구성

메시지의 형태		건수
일정이 포함되지 않음		3,564
일정이 포함됨	시간과 장소 모두가 포함됨	2,884
	장소만 포함됨	117
	시간만 포함됨	223
총 계		6,788

일정이 포함된 메시지 중에서 2,884건이고 나머지는 시간이나 장소만 포함된 메시지이다. 이하의 성능 평가를 위하여 이 말뭉치는 9:1의 비율로 학습 말뭉치와 실험 말뭉치로 분리해서 이용하였다.

성능 평가

성능 평가의 척도로는 일반적으로 정보검색에서 널리 사용되는 정확률(precision), 재현율(recall), F_1 점수(F_1 score)[24]를 사용하였다. 표 9는 띄어쓰기에 대한 성능평가 결과이다. 실험 말뭉치에 대해서 자동 띄어쓰기는 약 94%의 F_1를 보였으며 이 결과는 일반적인 자동 띄어쓰기 시스템[25]에 비해서 좋지 않지만, 단문 메시지의 여러 가지 특성을 고려할 때 그다지 나쁜 성능은 아니라고 판단된다.

표 10은 각각 띄어쓰기와 일정 추출에 대한 평가 결과를 보이고 있으며 표 11

표 9. 띄어쓰기에 대한 성능 평가

구 분	측정값
정답 말뭉치의 띄어쓰기 횟수(개)	5,006
자동 띄어쓰기 횟수(개)	5,098
정확히 일치한 띄어쓰기 횟수(개)	4,750
정확률(%)	93.17
재현율(%)	94.88
F_1 점수(%)	94.02

표 10. 일정(약속 시간/장소)에 대한 성능 평가

구 분	장소	시간	전체
정답 일정 수(개)	302	310	612
추출된 일정 수(개)	321	301	622
일치한 일정 수 (개)	257	269	526
정확률(%)	80.06	89.37	84.57
재현율(%)	85.10	86.77	85.95
F_1 점수(%)	82.50	88.05	85.25

은 자동 띄어쓰기를 했을 경우와 그렇지 않을 경우의 성능 차이를 보이고 있다. 표 10을 보면 약속 장소에 비해 약속 시간의 재현율과 정확률 모두 높게 나타났다. 이는 3.5절에서 기술한 장소 사전이 네이버 지도를 이용해서 수집되었기 때문에 단문 메시지에서 주로 나타나는 여러 다른 변이형을 충분히 포함하고 있지 못함을 간접적으로 보이고 있다. 또한 장소의 경우 말뭉치나 사전에 나타나지 않은 미등록어가 많고 시간에 비해 장소 뒤에 붙는 조사나 접사들이 생략되는 경우가 많기 때문으로 추측된다. 표 11에서 보듯이 띄어쓰기 정보가 일정을 추출하는데 중요한 자질임을 알 수 있다.

표 12은 메시지 단위의 정확률을 보이고 있는데 이 정확률은 메시지 내에 있는 모든 시간과 장소 모두를 추출할 비율을 의미하며 이 정확률은 일정이 포함된 메시지를 정확하게 찾아내는 비율이다. 즉 약 79%의 메시지에 대해서 정확하게 찾았고 이 메시지에 대해서는 추가적으로 일정을 입력할 필요가 없음을 말해준다.

표 11. 띄어쓰기 유무에 따른 성능 평가

구 분	장소	시간	전체
띄어쓰기를 한 경우 (F_1 점수(%))	82.50	88.05	85.25
띄어쓰기를 하지 않은 경우 (F_1 점수(%))	72.87	75.60	74.26

표 12. 메시지 단위의 정확률

구 분	전체
전체 메시지 수(개)	322
맞은 메시지 수(개)	253
정확률(%)	78.57

결 론

본 논문에서는 기계학습 기반의 단문 메시지로부터 일정(약속 시간 및 장소)을 추출하여 휴대 전화의 일정으로 자동 등록하는 시스템을 제안한다. 단문 메시지는 문법에 어긋나거나, 비속어나 신조어, 사투리 등을 많이 포함하고 있으므로 일반적인 자연언어 처리 모듈을 그대로 적용할 수 없다. 더구나 단문 메시지는 80 바이트로 자신의 의사는 전달해야 하므로 말을 줄여 쓰거나 띄어쓰기를 생략하는 경우가 자주 발생한다. 또한 단문 메시지를 대상으로 공개된 말뭉치가 없을 뿐 아니라 구축하는데도 많은 어려움이 존재한다. 이러한 제약에도 불구하고 본 논문에서 제안한 시스템은 85.25%의 F_1 점수와 78.57%의 메시지 단위 정확도를 나타내었고 실제 휴대전화에 탑재되어 실행하는데 아무런 문제가 발생하지 않았다. 그러나 상용화를 위해서는 인식률이 좀 더 개선되어야 한다. 이를 위해서는 말뭉치 부족 문제가 해결되고, 단문 메시지를 대상으로 하는 언어처리 엔진들이 개발된다면 더욱 높은 성능을 기대할 수 있을 것으로 예상된다.

참고문헌

[1] 최병목 외 (2005), **청소년의 휴대전화 사용실태 조사연구, 연구보고서** 05-08, 한국정보문화진흥원.

[2] J. Lee, H. Park, J. Ahn and M. Kim (1995), An Effective Indexing Methods for

Korean Text", *Proceedings of the Korean Society for Information Management Conference*, pp. 11-14, 1995.

[3] C. Jung (2004), *An Indexing Method Based on the Mixed n-gram for Korean Information Retrieval*, Master Thesis, Department of Computer Engineering, Korea Maritime University.

[4] 김재훈 (2004), **정보추출의 기술 현황**, 정보과학회지22-4, 35-46.

[5] http://www.chokkan.org/software/crfsuite/

[6] H. A. do Prado and E. Ferneda (2007), *Emerging Technologies of Text Mining: Techniques and Applications*, Idea Group Reference.

[7] C. Siefkes and P. Siniakov (2005), An Overview and Classification of Adaptive Approaches to Information Extraction. *Journal on Data Semantics* IV, *LNCS* 3730, pp. 172-212.

[8] A. L. Berger, S. A, Della Pietra, and V. J. Della Pietra (1996), A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, 22(1):1-36.

[9] V. N. Vapnik. (1998), *Statistical Learning Theory*, Wiley-Interscience.

[10] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel (1997), Nymble: A High-Performance Learning Name-Finder, *Proceedings of the Conference on Applied Natural Language Processing*, pp. 194-201.

[11] A. MaCallum, D. Freitag, and F. Pereira F. (2000), Maximum Entropy Markov Models for Information Extraction and Segmentation, Proceedings of the 17th International Conference on Machine Learning pp. 591-598.

[12] J. Lafferty, A. McCallum, and F. Pereira (2001), Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of International Conference on Machine Learning*, pp. 282-289.

[13] J. Darroch and D. Ratcliff (1972) Generalized Iterative Scaling for Loglinear Models, The Annals of Mathematical Statistics, 43:1470-1480.

[14] R. H. Byrd, J. Nocedal,, and R. B. Schnabel (1994), Representations of Quasi-Newton Matrices and Their Use in Limited Memory Methods, *Mathematical Programming*, 63(4):129-156.

[15] J. R. Finkel, T. Grenager, and C. Manning (2005) Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pp. 363-370, 2005.

[16] B. Wellner, A. McCallum, F. Feng, and M. Hay (2004), An Integrated, Conditional Model of Information Extraction and Coreference with Application to Citation Matching, Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 593-601.

[17] X. Zhang and A. C. Fang (2010) Term Recognition Using Conditional Random Fields, Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering.

[18] http://crfpp.sourceforge.net/

[19] http://www.chokkan.org/software/crfsuite/benchmark.html

[20] http://www.microsoft.com/korea/smserver/default.mspx

[21] http://www.iconsoft.com/products/phonex/

[22] L. A. Ramshaw and M. P. Marcus (1995), Text Chunking Using Transformation-Based Learning, Proceedings of the Third Workshop on Very Large Corpora, pp. 82-94.

[23] http://map.naver.com/

[24] 안동언, 김재훈, 남영준, 박혁로, 이상곤 (2010), **최신정보검색론**, 교보문고.

[25] 이도길, 이상주, 임희석, 임해창 (2003), 한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델, 정보과학회 논문지: **소프트웨어 및 응용, 30**(4):358-371.

1 차원고접수 : 2010. 10. 20
2 차원고접수 : 2011. 1. 18
최종게재승인 : 2011. 1. 21

(*Abstract*)

Automatically Registering Schedules from SMS Messages on Handheld Devices

Jae-Hoon Kim Hyung-Chul Kim

Department of Computer Engineering, Korea Maritime University

With rapid spread of handheld devices like cellular or smart phones, a short message service (SMS) comes on the public as a communication means. SMS is very cheap and can be easily written down on the storage in order not to forget it, hence it is widely used to inform schedules (time and place). In this paper, we develop a system for automatically registering schedules extracted from SMS text messages. SMS text messages are very short and concise, but include a lot of Internet words like slangs and abbreviations. These have made it difficult to extract information on schedules from them. Also handheld devices have some limitations on computing power and storage and then applying general natural language processing modules like morphological analysis to the devices are somewhat hard. To relax these burdens, we extract schedule informations from SMS messages using machine learning methods like condition random field (CRF) without using any language processing modules and register the informations on the schedule management system of handheld devices. Our proposed automatic schedule registration system has implemented on Samsung Omnia phone for experiments.

Key words : SMS messages, Schedule Information Extraction, Schedule Management of Handheld Devices