인지과학, 제21권 제4호
Korean Journal of Cognitive Science
2010, Vol. 21, No. 4, 447~458.

시간의 단위별 처리를 이용한 자동화된 한국어 시간 표현 인식 및 정규화 시스템*

선 충 녕　　강 상 우†　　서 정 연
서강대학교 컴퓨터학과

시간 정보는 문서나 문장 등에서 매우 중요한 정보로 사용되기 때문에 다양한 종류의 데이터에서 시간 정보의 인식은 매우 중요하다. 시간 정보는 일정한 형태를 가진 것으로 간주되지만 실제 사용되는 시간 표현은 매우 다양하고 복잡하며 정보의 일부가 빈번하게 생략되는 경우가 발생한다. 본 연구에서는 시간 표현의 추출뿐만 아니라 추출된 표현을 정규화된 표준 형식으로 변환하는 범용 시간 표현 추출 및 변환 시스템을 제안한다. 다양한 시간 표현의 추출과 변환에 필요한 노력을 줄이고 새로운 데이터에 대한 확장성을 보장하기 위해 기본 시간 단위를 정의하였다. 추출단계에서는 기본 시간 단위의 조합으로 구성된 사전을 사용하여 가능한 시간 표현들을 추출한다. 정규화 변환 단계에서는 인접 추출 정보와 기준 시간 등을 사용하여 생략된 기본 시간 단위 정보를 복원하고 최종적으로 모든 기본 시간 정보들은 통합되어 정규화된 표준 형식으로 변환된다. 제안한 시스템은 모바일 기기 등의 잡음 환경에서 강인한 성능을 보장하며 영역이나 언어에 대해 독립적이므로 많은 영역에서 응용이 가능하다. 본 연구는 실험에서 다량의 오류가 포함된 SMS 데이터에서 시간 표현 추출 정확도 93.8%, 시간 표현 변환 정확율 93.2%을 보임으로써 오류에 강인하면서도 높은 성능을 유지함을 증명하였다.

주제어 : 정보 추출, 시간 관계, 모바일 인터페이스, 인간-기계 상호작용, 자연어처리, 유비쿼터스

---

* 이 논문 또는 저서는 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2010-0027-797)

† 교신저자: 강상우, 서강대학교 컴퓨터학과, 연구 분야: 자연어처리,음성대화시스템

서 론

시간 정보는 정보 추출, 질의 응답 시스템, 자동 요약 등 자연어처리의 여러 응용 분야에서 중요한 역할을 한다. 기존의 문서나 기사 등에서의 시간 표현은 일정한 형식을 가지고 있기 때문에 수동 구축된 규칙으로 손쉽게 추출 및 변환이 가능했다. 그러나 최근 웹과 모바일 기기의 활용이 늘어나면서 다양하고 복잡한 형태의 시간 표현들을 효과적으로 추출할 수 있는 방법이 요구되고 있다.

시간 표현을 추출하기 위한 전통적인 접근 방법은 시간 표현이 정형화된 형태로 나타날 것을 가정한다. 따라서 시간 표현을 추출하기 위해 최적화된 규칙을 수동으로 작성하고 추출된 결과를 정규화하기 위한 모듈을 구현하는 것이 일반적이었다. 하지만 수동 규칙을 이용하는 방법은 시간 표현의 형태가 다양하며 복잡한 표현 형태를 갖는 경우에는 규칙의 적용이 용이하지 않다. 또한 기존 규칙을 새로운 영역에 적용하기 위해서는 시간 표현을 수집한 후 전문가에 의해 분석 및 규칙의 재작성이 필요하기 때문에 이식성이 떨어진다. 뿐만 아니라 단문 메시지(Short Message)에서와 같이 참여자들이 시간 표현을 빈번하게 생략하는 경우에는 규칙의 작성이나 정규화 과정이 더욱 복잡하고 어렵게 된다.

본 연구는 문장에 포함된 시간 표현을 추출하고 추출된 시간 표현들을 정규화된 표현으로 변환함으로써 여러 영역에 범용적으로 적용 가능한 방법을 제안한다. 시간 표현을 각각 기본이 되는 단위(Atomic Unit)로 분해하여 추출과 정규화 과정을 거치고, 기준 시간(Basis Time)과 인접 추출 정보를 이용하여 생략된 시간 정보를 복원한다, 최종적으로 기본 시간 정보들을 통합하고 정규화된 표준 형식으로 변환한다. 제안 방법은 변형되지 않는 시간의 최소 단위를 대상으로 규칙을 작성하고 이들의 조합을 통해 시간 표현을 인식하고 변환한다. 따라서 새로운 시간 표현이나 영역의 변화에도 기본 단위의 조합을 통해 대부분의 시간 표현이 처리될 수 있으므로 높은 확장성을 가진다. 또한 외부적으로 제공되는 기준시간과 추출된 기본 시간 단위의 조합을 통해 생략된 시간 표현이나 상대적인 시간 표현도 효과적으로 처리할 수 있다.

관련 연구

자연어처리 분야에서 태그 부착 말뭉치의 구축은 기술의 개발과 실험의 성과를 이루는데 매우 중요하다. 이와 관련하여 다양한 문서에서 필요한 정보의 추출과 표준화에 대한 많은 연구들이 진행되고 있으며 Defense Advanced Research Projects Agency(DARPA) Translingual Information Detection Extraction, and Summarization(TIDES)과 Automatic Content Extraction(ACE). 연구 프로그램의 지원을 통하여 많은 성과를 이루어 졌다[1][2][3][4][5][6]. 질문-대답 관계(Question Answering Relation), 사건 특징화와 추적(Event Characterization and Tracking), 시간상의 사건 시각화(Visualization of Events on Timelines), 약력 생성(Production of Biographical Summaries)등의 응용분야 에서 그 성과들은 검증되었다. 기존 연구가 신문기사와 같이 정형화된 형태의 문서들을 대상으로 했다면, 최근에는 웹문서와 단문메시지 등 다양하고 복잡한 시간 표현을 포함하는 문서들로 그 대상이 확대되고 있다. 시간 표현의 추출 연구는 DARPA에서 지원한 Message Understanding Conference 7(MUC7)에서 활발히 연구가 진행되었다. MUC7에서 시간 표현의 추출은 표현이 보다 명확히 표기된 영어 신문으로부터 문법패턴을 사용하여 추출하고 효과적인 문법패턴의 개발을 위하여 많은 연구들이 진행되었다. 기존 연구들은 미리 인식하고자 하는 패턴을 등록하고 이 패턴에 의해 시간 표현을 추출하였다[7][8]. 또한 표준 형식으로 변환하는 연구 역시 수동 작성된 규칙을 기초하여 기계 학습을 반영하는 정도의 연구가 진행되고 있다[2]. 수동 규칙에 기초한 접근 방법은 시간 표현마다 각각 인식 및 변환을 위한 규칙이나 오토마타를 구현해야 하기 때문에 다른 영역에서는 새롭게 작업을 하는 경우가 대부분이었다. 또한 추가적인 시간 표현을 등록하기 위해서도 필요한 패턴을 분석하고, 이에 대한 변환 모듈을 다시 구현하는 과정이 필요하다. 이에 비해 본 연구에서는 인식 및 변환 규칙을 기본 단위로 분할하여 구축하기 때문에 조합을 통해 새로운 패턴의 인식과 변환이 가능하다. 또한 자동으로 추출 패턴을 조합하는 추출 패턴 생성 단계를 이용하여 수동으로 규칙을 생성하는 시간을 최소화 하였다.

제안 시스템 구성

본 연구에서는 시간 표현을 시간의 기본 단위로 분리하여 각각 인식하며, 인식된 결과를 조합하는 방법으로 적은 인식기를 가지고 많은 표현들을 처리할 수 있다. 또한 기준 시간과 인접 추출 정보를 이용하여 생략된 단위 시간 정보를 복원하고 기계가 이해할 수 있는 정규화된 형태로 출력을 제공할 수 있다. (그림 1)에서 보는 것과 같이 제안 시스템은 시간표현 추출 패턴 생성 과정과 시간 정보 추출 과정의 두 부분으로 구성된다. 시간 표현 추출 패턴 생성 과정은 정답 문자열 파일을 분석하여 시간 표현 추출 및 정규화에 필요한 정보를 구축하는 단계이다. 시간 정보 추출 과정은 추출 패턴DB를 이용하여 시간 표현과 같은 표현의 문자열들을 추출하고 컴퓨터가 인식할 수 있는 정규화 표현으로 변환한다.

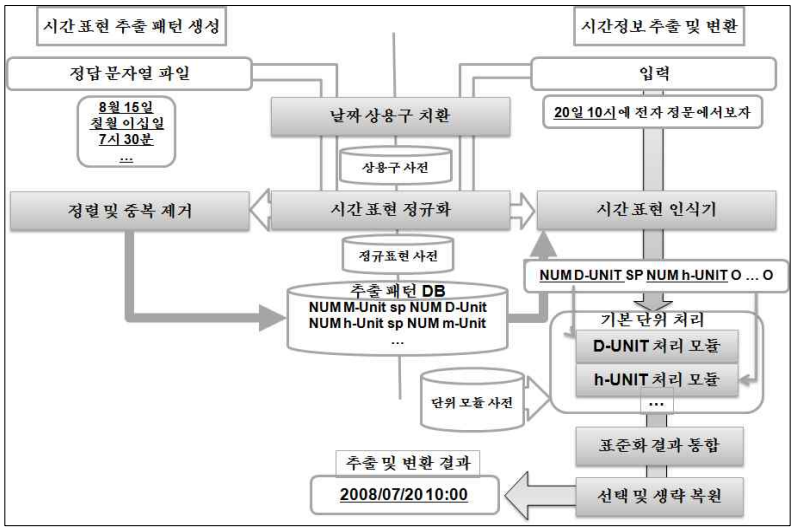

그림 1. 전체 시스템 구성 및 동작 예

영역 시간 표현을 이용한 추출 규칙 자동 생성

시간은 하나의 의미를 위해 매우 다양한 표현들이 존재한다. 각각의 표층적인

표현마다 개별적으로 처리하기 보다는 비슷한 의미를 나타내는 정보들을 결합하여 함께 처리하는 것이 효율적이다. 본 연구에서는 시간의 표층적인 표현에 대해 처리하는 것이 아니라 각각의 정규화된 심볼 단위로 처리한다. 다양한 표현 형태를 가지더라도 같은 유형의 심볼 열로 변환된다면, 기존의 추출 및 변환 과정을 그대로 사용할 수 있기 때문에 시스템은 각 단위 처리기의 수가 극적으로 증가하는 것을 막을 수 있다.

제안 시스템에서는 대상 영역의 시간 표현 열들을 모은 정답 문자열로부터 각각을 추출할 수 있는 의미 심볼 열로 치환한다. 치환된 의미 심볼 열은 추출패턴 DB에 등록되어 문장 내에 포함된 시간 표현을 추출할 수 있는 시간 표현 추출 패턴으로 사용된다.

새로운 시간 표현이나, 영역에 따라 다른 시간 표현들은 새롭게 재작성된 시간 표현이라기보다는 기존의 시간 표현이 재조합된 경우가 대부분이다. 따라서 대부분의 새로운 시간 표현들은 정답 문자열에 등록하는 것만으로도 추출 및 변환을 위한 규칙을 생성할 수 있다. (그림 2)의 왼쪽 상단의 상자는 영역에서 수집된 정답 표현들을 보여주며, 오른편은 그 사전을 정규 표현으로 변환하여 작성된 추출 패턴들을 보여준다.

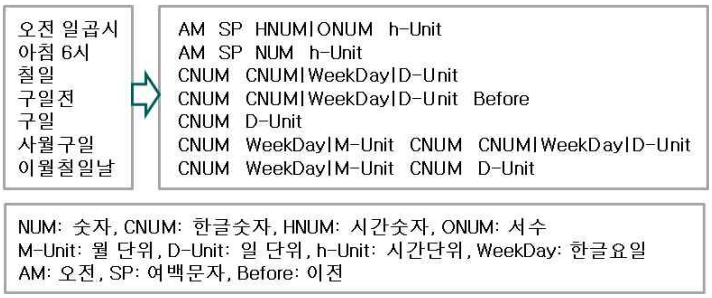

그림 2. 추출 패턴 DB 예

시간 정보 추출

시간 정보 추출 단계에서는 입력을 날짜 상용구 치환과 시간 표현 정규화 과정

을 통해 변형시킨다. 이후, 변환된 입력에 포함된 시간 표현들은 추출 패턴 DB에 등록된 정규화 패턴에 의해 추출된다. 이때 가능한 모든 후보들이 추출되며, 중복되는 경우에는 더 긴 시간 표현을 결과로 선택한다. 시간 정보 추출에서 정보를 가지고 있지 않는 국경일과 같은 표현들은 일반적인 방법으로 처리될 수 없다. 본 연구에서는 이와 같은 날짜를 나타내는 상용 표현들이 의미하는 날짜를 상용구 사전에 등록하는 방법으로 해결하였다. 이를 통해 새로운 상용구의 추가가 필요할 경우 새로운 오토마타 작성 등의 추가 작업 없이 상용구 사전에 새로운 문자열을 추가하는 것만으로 시간 정보를 추출할 수 있다.

하나의 시간 표현들은 년, 월, 일, 시, 분, 초 단위로 구성된다. 실제로 사용하는 시간 표현들은 각각의 단위에 따라 다양하게 조합되어 사용될 수 있기 때문에 표층적인 패턴마다 정규화하는 것은 사실상 불가능하다. 따라서 본 연구에서는 하나의 시간 표현을 하나의 처리 단위로 분석 하지 않고 시간의 기본 단위를 분석의 단위로 사용한다. 이 단계에서는 인식된 시간 표현들을 시간의 기본 단위들로 나누고 각각을 기계가 인식할 수 있는 형태로 표준화하는 작업을 수행한다. 이 때, 상대적인 표현들은 각 단위에 대한 명시적인 시간 값이 들어가는 것이 아니라 특정 기준 시간에 대한 상대적인 값으로 변환한다.

표준화 결과 통합단계는 단위 표준화 모듈에서 시간 기본 단위로 나누어진 정보들을 하나의 정보로 통합한다(그림 3). 우선, 다른 시간 추출 결과와 겹치지 않는 시간 표현 중에서 충분한 시간 정보를 가지는 절대적인 시간 표현이 먼저 통합된다. 이렇게 통합된 정보를 사용하여 주변의 상대적인 시간 표현들은 부족한 정보를 보충한다. 만약 주변에 절대 시간 표현이 없는 경우, 시스템 및 문서 단위에서 제공되는 시간을 이용하여 상대시간을 절대 시간으로 변경하는 작업이 수행된다. 만일 시간 표현이 중첩되는 경우에는 영향을 미치는 시간 표현들 중에서 선택해야 하는 경우에는 더 자세한 정보를 제공하는 시간 표현이 선택된다. 최종적으로 추출된 시간 표현들은 유효한 시간 범위, (1~12월, 0~24시, 0~60분 등), 에 해당하는 값들을 가지고 있는지를 검토하여 결과를 통합한다.

일상적인 대화에서는 이전에 출현하지 않은 시간에 대한 정보라도 습관적으로 사용하는 정보의 생략이 발생한다. 예를 들어, “2시에 보자”와 같은 문장에서의 ‘2시’는 새벽 2시가 아닌 오후 2시인 14시를 의미한다. 하지만 이것은 상식적인 정보

를 생략하기 때문에 시간 정보를 추출하는 시스템은 이와 같은 시간 및 날짜 표현을 처리할 수 있어야 한다. 본 연구에서는 이와 같은 시간 표현을 위해 설정 파일을 통해 우선순위를 지정하고 있다. 하지만 이러한 표현은 생활의 패턴에 따라 기준이 달라질 수 있기 때문에 설정을 통해 상황에 맞춰 사용할 수 있도록 구성하였다. 상대적인 시간이나 시간 정보의 일부가 생략된 경우, 시스템은 추출된 정보 주변에서 나타나는 절대 시간이 가지는 정보를 이용하고 만약 이용할 수 있는 절대 시간이 존재하지 않는다면 미리 설정된 기준 시간에 의해 생략된 정보를 보충한다.

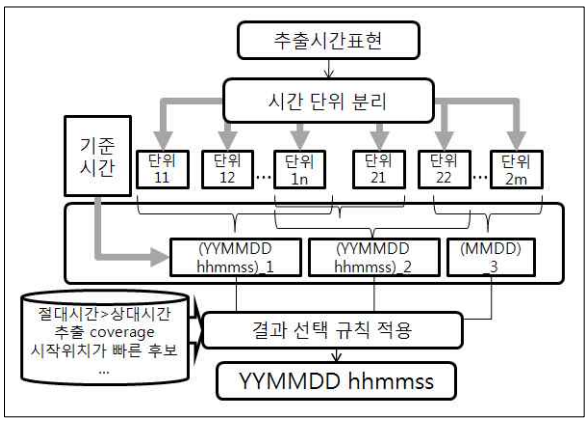

그림 3. 표준화 결과 통합 모듈의 동작 예

실험 및 결과

제안 시스템을 검증하기 위해 SMS 데이터 말뭉치를 사용하였다. SMS 데이터는 비교적 정형화된 신문 등의 문서와는 달리 실제 메시징 서비스를 사용하면서 발생하는 띄어쓰기 오류, 철자오류, 비속어 등의 오류가 포함되므로 보다 실제 모바일 환경과 유사한 상황에서 실험이 가능하다. 본 절에서는 제안 시스템의 성능 측정과 분석 결과를 설명하고 오류의 원인과 해결 방안에 대하여 논의한다.

실험 환경

본 실험에서 사용된 코퍼스는 약속에 관련된 SMS 데이터이며 80명의 대학생 참가자들에 의해 실제 휴대폰을 사용하여 수집되었다. 시간에 관련된 코퍼스를 위해 수집된 전체 6,170개의 데이터 중에서 약속을 정하는 내용과 관련된 데이터 4,685개를 선정하여 구성하였다. 이 데이터는 관련 내용을 교육 받은 두 명의 대학원생에 의해 수집된 SMS에서 시간 및 날짜에 관련된 내용을 정보를 부착하고 교차 검토를 수행하였다. 수집된 데이터는 3,670개의 날짜 정보와 4,516개의 시간 정보를 포함하고 있다. 임의로 추출한 333개의 SMS 문장을 성능 측정을 위해 선택하였고 나머지 4,352 문장을 학습 데이터로 사용하였다.

실험 결과

말뭉치의 각 문장에 포함된 시간 표현에 대해 추출 성능을 측정하고 추출된 부분을 정규화된 표준 형식으로 바꾸는 변환 성능을 별도로 측정하였다. (표 1)은 제안 시스템의 시간 표현 추출과 변환 성능을 보여준다. 전체 말뭉치 333개 중에 포함된 시간 표현의 개수는 353개이며 제안 시스템은 시간 표현 추출에서 93.77% 의 정확률을 보여준다. 또한 추출된 시간 표현에 대한 시간 표현 정규화 변환 정확률은 93.20% 이다. 정규화 성능이 추출성능보다 다소 낮은 수치를 보이는 이유는 추출 과정에서의 오류가 정규화 과정으로 전파되어 정규화 성능에 영향을 미치기 때문이다. 따라서 시간 표현 추출이 정확한 경우에 대해서만 정규화 변환 정확률을 측정한 경우에는 전체 데이터의 정확률보다 높은 95.17%의 정확률을 보인다.

표 1. 시간 표현 추출 및 정규화 성능

	정답 수	오답 수	정확률
시간 표현 추출 성능	331	22	93.77%
시간 표현 정규화 성능	329	24	93.20%

(표 2)는 시간 표현 추출 과정에서 잘못 추출된 문장의 예를 보여준다. 1, 2번 문장은 띄어쓰기와 철자오류로 인해 시간 표현 추출이 실패한 경우이다. 1번 문장에서는 “내 일”에서 띄어쓰기에 오류가 있었고 2번 문장에서는 “모래”에서 철자오류가 발생하여 추출에 실패한 경우이다. 3번과 4번 문장의 경우에는 증거 단어로 추출된 단어들이 다른 내용어들의 일부분인 경우이다. 3번 문장에서 추출된 “스승의날”은 명시적인 날짜가 아니고 “스승의날 기념”의 일부로 보아야 할 것이다. 4번 문장 또한 “저녁”은 특정 시간 구간을 의미하는 것이 아니고 “저녁 먹으러”의 일부 단어로서 “식사”의 의미를 갖는 것으로 보는 것이 타당하다. 이런 경우는 주변단어들과의 관계를 고려해야 하여 추출대상에서 제외할 수 있는 추가 연구가 수행되어야 할 것 이다. 5번 문장은 문장 내용의 의미해석이 필요한 경우이다. 추출된 “두시”는 명시적인 시간의 의미가 아니라 현재부터 “두 시간 후”의 의미를 갖는다. 이 경우 패턴 추가에 의해서 일부 수정이 가능하지만 근본적인 성능 향상을 위해서는 복잡한 의미 해석 과정이 필요하다.

표 2. 시간 표현 추출 오류 예

번호	SMS 문장	시간 추출
1	인석아 우리 내일 한 시에 수영장갈껀데 같이 가자	한시
2	웃찾사 관람을 원하시는 분은 내일 모래 여덟시까지 방송국 앞으로 와주십시오	내일_여덟시
3	5월14일 스승의 날 기념으로 대원여고 앞에서 3시에 보자 ㅋㅋㅋ	5월14일, 스승의날_3시
4	동생학원 끝나고 저녁 먹으러 가게 학교 끝나고 바로 집으로와	저녁
5	지금 니네 집 가고 있다 두 시간쯤 걸릴 거 같은데 시간 맞춰서 집 앞에 나와 있어 간장게장 주게	두시

(표 3)은 시간 표현 정규화 과정에서 발생한 오류의 예를 보여준다. 1번 문장은 추출된 “금날 7시”가 “오늘 7시”의 정규화 표현인 “20080807190000”로 변환된 경우이다. “금날”은 철자오류 혹은 신조어로 분류할 수 있으며 등록된 유사 패턴 중

(“금일”, “금(요일)”)에서 모호성일 발생하여 “금일”로 잘못 변환된 경우로 해석할 수 있다. 2번 문장은 추출된 “내일”이 문장의 마지막으로 도치되어 추출된 “6시부터”가 기준 날짜인 오늘로 변환되어 정규화 표현이 “20080807180000”로 생성된 경우이다. 3번 문장은 의미적으로 오류가 포함된 문장으로 “오늘”, “새벽 2시”의 조합은 이미 지난 시간으로 “20080807020000”이 아니라 하루 뒤인 “20080808020000”로 변환 되는 것이 타당하다. 변환 오류의 대부분의 문제는 철자오류 혹은 신조어 문제로서 빈번하게 사용하는 단어들을 추가함으로써 해결이 가능하다. 2번 문장과 같은 오류는 구문 분석 정보와 같은 전통적인 언어처리 기법이 요구된다. 하지만 본 연구의 대상은 기존의 언어처리 기법의 대상과 달리 많은 오류를 포함하기 때문에 오류에 강인한 언어처리 기법 연구가 선행되어야 할 것이다. 마지막 오류 형태는 사용자의 표현이 의미적으로 오류를 포함하기 때문에 기존 언어처리 기법을 적용하여 보다 많은 정보를 분석할 수 있는 경우에서도 해결하기 어렵다. 이 경우에는 시스템이 의미적 모순을 파악해야 있어야 하기 때문에 새로운 접근의 시도가 필요할 것으로 생각된다.

결 론

본 논문에서는 단위 시간의 추출과 통합에 의해 시간 표현 추출과 생략 복원 그리고 정규화 과정을 포함하는 범용 시간 표현 추출 및 변환 시스템을 제안하였다. 제안한 방법은 새로운 영역에서 시간 표현을 추출하고 기존 모델을 수정하기 위한 시간과 노력을 줄일 수 있으며 추출 대상을 추가와 모델의 확장이 용이하기 때문에 광범위한 응용이 가능하다. 뿐만 아니라 시간 표현 추출을 위한 지식 구축 작업 및 변환 과정이 특정 언어에 의존적인 형식이나 지식을 사용하지 않기 때문에 구축된 지식의 수정만으로 한국어뿐 아니라 다양한 언어에도 활용될 수 있는 범용 시간 표현 추출 및 변환 시스템으로 사용될 수 있다.

참고문헌

[1] Inderjeet Mani, George Wilson, "Temporal Granularity and Temporal Tagging of Text," In *Bettini C., Montanari A. (eds.),* In *Proceedings of the AAAI Workshop on Spatial and Temporal Granularity*, pp.71-73, 2000.

[2] Inderjeet Mani, George Wilson, "Robust temporal processing of news," In *Proceedings of the 38th Association for Computational Linguistics*, pp.69-76, 2000

[3] Wilson, George, Interjeet Mani, Beth Sundheim and Lisa Ferro, "A multilingual approach to annotating and extracting temporal information," In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, pp.81-87, 2001

[4] Inderjeet Mani, George Wilson, Lisa Ferro and Beth Sundheim. "Guidelines for annotating temporal information," In *Proceedings of the first international conference on Human language technology research*, pp.1-3, 2001.

[5] laurie Gerber, Lisa Ferro, Inderjeet Mani, Beth Sundheim, George Wilson and Robyn Kozierok, "Annotating temporal information: from theory to practice," In *Proceedings of the 2002 Conference on Human Language Technology*, pp.226-230, 2002.

[6] Wiebe, Janyce M., Thomas P. O'Hara, Thorsten Ohrstrom-Sandgren and Kenneth McKeever, "An empirical approach to temporal reference resolution," *Journal of Artificial Intelligence*, Vol.9, pp.247-293, 1998.

[7] Andrei Mikheev, Claire Grover and Marc Moens, "Description of the LTG system used for MUC-7," In *Proceedings of the 14th Message Understanding Conference (MUC-7)*, 1998.

[8] Rohini Srihari, Cheng Niu and Wei Li, "A hybrid approach for named entity and sub-type tagging," In *Proceedings of the sixth conference on Applied natural language processing*, pp.247-254, 2000.

1 차원고접수 : 2010. 2. 19
2 차원고접수 : 2010. 9. 1
최종게재승인 : 2010. 10. 6

(*Abstract*)

Automatic Recognition and Normalization System of Korean Time Expression using the individual time units

Choong-nyoung Seon Sangwoo Kang Jungyun Seo

Department of Computer Science, Sogang University

Time expressions are a very important form of information in different types of data. Thus, the recognition of a time expression is an important factor in the field of information extraction. However, most previously designed systems consider only a specific domain, because time expressions do not have a regular form and frequently include different ellipsis phenomena. We present a two-level recognition method consisting of extraction and transformation phases to achieve generality and portability. In the extraction phase, time expressions are extracted by atomic time units for extensibility. Then, in the transformation phase, omitted information is restored using basis time and prior knowledge. Finally, every complete atomic time unit is transformed into a normalized form. The proposed system can be used as a general-purpose system, because it has a language- and domain-independent architecture. In addition, this system performs robustly in noisy data like SMS data, which include various errors. For SMS data, the accuracies of time-expression extraction and time-expression normalization by using the proposed system are 93.8% and 93.2%, respectively. On the basis of these experimental results, we conclude that the proposed system shows high performance in noisy data.

Keywords : information extraction, temporal relation, mobile interface, human-machine interaction, NLP, ubiquitous