사건 탐지 및 추적을 위해 신문기사에서 자동 추출된 시간정보의 유용성 판단

(Judgment about the Usefulness of Automatically Extracted Temporal Information from News Articles for Event Detection and Tracking)

김 평 † 맹 성 현 ††
(Pyung Kim) (Sunghyon Myaeng)

요 약 시간정보는 정보 추출, 질의응답 시스템, 자동 요약과 같은 자연언어 처리 응용분야에서 중요한 역할을 한다. 사건 탐지 및 추적 분야에서는 기사의 발행일이 기사간 유사도 계산에 많이 사용되고 있지만 그 유용성에는 한계가 있다. 본 연구에서는 사건 탐지 및 추적 시스템의 성능을 향상시키기 위해서, 한국어 신문기사를 대상으로 비교적 간단한 자연언어 처리 기술을 사용해서 시간정보를 추출하는 방법을 개발하였다. 시간표현 어구를 추출하기 위해 품사패턴과 어휘사전이 사용되었고, 추출된 시간표현 어구는 정규화 과정을 통해 특정 시각 또는 기간으로 변환되었다. 실험을 통해 시간표현 추출과정의 정확도를 측정하였고, 기사에서 자동으로 추출된 시간을 사용함으로써 사건 탐지 및 추적 시스템의 성능을 향상시킬 수 있었다.

키워드 : 시간정보 추출, 사건 탐지 및 추적

Abstract Temporal information plays an important role in natural language processing (NLP) applications such as information extraction, discourse analysis, automatic summarization, and question-answering. In the topic detection and tracking (TDT) area, the temporal information often used is the publication date of a message, which is readily available but limited in its usefulness. We developed a relatively simple NLP method of extracting temporal information from Korean news articles, with the goal of improving performance of TDT tasks. To extract temporal information, we make use of finite state automata and a lexicon containing time-revealing vocabulary. Extracted information is converted into a canonicalized representation of a time point or a time duration. We first evaluated the extraction and canonicalization methods for their accuracy and investigated on the extent to which temporal information extracted as such can help TDT tasks. The experimental results show that time information extracted from text indeed helps improve both precision and recall significantly.

Key words : temporal information extracting, event detection and tracking

1. 서 론

시간정보는 자연어 처리의 응용분야는 물론 정보검색의 응용에도 매우 중요한 역할을 수행하는데, 질의응답 시스템에서는 '언제'에 대한 답변으로, 정보추출 시스템에서는 시간정보 템플릿(예 : 사건이 발생한 시간)을 채우기 위해서 사용되었다. 이런 시스템들에서는 시간정보가 추출될 최종 목표중 하나이지만, 최근 연구되고 있는 시간 기반 자동 요약 분야[1](같은 주제를 다루는 다수의 뉴스 기사로부터 시간정보를 사용하여 주제문을 선택)에서는 성능 향상을 위한 하나의 도구로서 사용되고 있다.

사건 탐지 및 추적(Topic Detection and Tracking : TDT)은 시간정보가 매우 유용하게 사용되는 분야중 하나이다[2-7]. 사건 탐지는 사건별로 기사들을 구분하거나 또는 온라인 기사에서 새로운 사건을 찾는 과정이고, 사건 추적은 특정 사건에 대한 단서 문서를 가지고 동일 사건에 대해 다룬 기사들을 찾는 과정이다. 여기서

† 정 회 원 : 한국과학기술정보연구원 NTIS 사업단 연구원
pyung@kisti.re.kr
†† 종신회원 : 한국정보통신대학교 공학부 교수
myaeng@icu.ac.kr
논문접수 : 2004년 4월 9일
심사완료 : 2006년 4월 26일

사건은 "특정 시간과 장소에 발생한 어떤 일"을 의미한다[2]. 즉 시간과 장소는 두 개의 기사가 동일 사건에 대해 보도하는지 또는 비슷한 다른 사건에 대해 보도하는지를 결정하는 중요한 요소로 동작한다. 따라서 기사에서 보도하는 사건에 대한 정확한 시간 추출은 시간순으로 기사 또는 사건 관련 문장들을 정렬함으로써 사건 탐지 및 추적에 중요한 역할을 할 수 있다.

특정 사건에 대한 기사는 일정기간에 보도되고 그 기간이 경과한 후의 기사는 새로운 사건에 대한 기사일 가능성이 높다는 가정하에, 대부분의 TDT 시스템들은 발행일 순으로 정렬된 기사집합에서 일정기간에 속하는 기사들을 동일 사건에 대한 후보로서 사용하였다. 즉 기사의 발행일에 따른 시간 차이를 가중치에 반영하여 같은 사건에 대한 기사 여부를 판별하는 기준으로 사용하였다[2-6]. 그러나 항상 기사가 최근에 발생한 사건만을 보도하는 것은 아니며, 또한 사건의 전개에 따라 기사가 시간 차이를 두고 보도되는 경우 동일 사건에 대한 기사임에도 불구하고 서로 다른 사건으로 판별될 수 있다.

본 연구에서는 한국어 신문기사를 대상으로 기사에 나타난 절대시간은 물론 상대시간을 자동으로 추출하고, 이를 기사의 발행일을 기준으로 특정 시각이나 기간으로 정규화하였다. 이렇게 추출된 시간정보는 동일 사건을 보도하는 기사들을 보다 정확하게 찾아내기 위해 사용되었다. 실험을 통해 시간정보 추출과정의 정확도는 물론 기사에서 자동 추출된 시간이 TDT 시스템의 성능에 미치는 영향을 알아보았다.

2장에서는 시간정보 추출을 위한 기존 연구와 TDT 시스템에서 시간이 사용되는 방법에 대해 기술하고, 3장에서는 본 연구에서 제시하는 시간 추출 및 정규화 방법을 기술한다. 4장에서는 실험을 통해 시간 추출 시스템의 단계별 정확도, 추출된 시간이 TDT 시스템의 성능에 미치는 영향에 대해 기술한다. 마지막으로 5장은 연구 결과와 향후 연구 방향에 대해 기술한다.

2. 관련연구

정보추출(IE) 분야에서 문서로부터 구조 템플릿을 채우기 위한 하나의 방법으로 시간정보를 추출하는 것과 관련된 연구가 진행되기 시작했다. MUC(Message Understanding Conferences)-6에서는 개체명 인식의 하부작업으로써 절대시간을 구분짓는 연구가 이루어졌고[8], MUC-7에서는 이를 확장하여 상대시간까지도 개체명에 포함시켰다[9]. 그렇지만 MUC에서는 시간정보와 관련된 연구는 제한적이었으며 성능도 좋지 않았다.

미국 DARPA의 TIDES(Translingual information Detection, Extraction, and Summarization) 프로젝트에서는 시간표현 지침(Temporal Guidelines)을 통해 시간정보 표현을 ISO8601 형식으로부터 유도된 그레고리안 캘린더(Gregorian calendar) 방식에 기초를 두고 'YYMMDDhhmmss' 형식으로 시간을 표시하도록 하였다. 시간표현을 위한 방법은 어느 특정 시점과 범위나 특정 시점을 기준으로 한 기간표현으로 나타내었다[10].

시간을 추출하기 위한 방법으로 형식담화(formal discourse) 기반이나 말뭉치(corpus) 기반에 의거하여 문서로부터 시간정보를 추출하기 위한 방법들이 큰 주류를 이루고 있다. 형식담화 기반 접근방법[11,12]은 자연어에 나타나는 표현 방식을 이해하고 의미 체계를 연구하는 방법으로, 특히 문맥에 나타난 시간정보를 바탕으로 사건의 흐름을 추론할 수 있게 한다. 실제 우리가 사용하고 있는 담화에서는 시간어구가 명확히 드러나는 경우보다는 내포적으로 나타나 있는 경우가 많다. 그러므로 자연어 문장의 구조를 분석하고 문맥을 이해해 내재된 의미를 추론하는 담화 분석은 시간정보를 추출하는데 사용할 수 있다. 말뭉치기반 접근 방법[13]으로 말뭉치로부터 구축된 어휘정보를 이용하여 문서내의 시간정보를 찾아내고, 그 시간정보의 문법적 역할을 활용하여 구문 분석시 발생하는 모호성을 제거하고자 하는 방법이 연구되었다. 이 연구에서는 용어 색인기를 이용하여 시간표현에 사용되는 단어를 추출하고, 의미와 기능에 따라 범주화 한 후, 시간 어구와 일반명사 어구간의 공기 정보를 통해 이들이 어떻게 하나의 의미를 구성하는지 설명했다. 이렇게 시간 어구와 일반명사 어구가 복합적으로 이루어진 시간표현은 유한오토마타(FSA)를 통해 유효한 시간표현으로 인식되고, 최종적으로 문장 내에서 해당 시간표현의 문법적인 역할이 무엇인지를 알아낼 수 있다.

TDT에 대한 연구는 사건을 기반으로 기사들을 구분하기 위해서 다음과 같은 다섯가지의 작업들로 구분되어 진행되고 있다[14].

- 기사 분할 : 방송기사의 사본을 의미적으로 연관된 각각의 기사들로 분할
- 새로운 사건 탐지 : 온라인 뉴스 기사들로부터 새로운 사건 여부를 탐지
- 사건 탐지 : 기사들을 사건별로 할당하는 작업
- 사건 추적 : 주어진 단서 기사를 가지고 동일 사건을 보도하는 기사들을 추적
- 기사 링크 탐지 : 두 개의 기사가 서로 같은 사건을 보도하고 있는지 여부를 판단

클러스터링 알고리즘을 사용해서 새로운 사건에 대한 기사를 탐지하고 사건별 기사를 군집시키는 연구[2-4,15], 정보 필터링 방법을 사용해서 사건을 추적하는 연구[2], 사건 기술에 사용된 각 어휘들간의 상호 정보를 사용한 어휘 체인을 사용한 연구[15]와 사건 또는

주제와 관련된 용어들을 구분하고 사건 관련 용어를 대상으로 사건을 찾는 연구[16]등 다양한 언어학적 기술이 사건 탐지 및 추적에 사용되고 있다. 본 연구는 기존의 TDT 시스템들이 기사의 발행일을 유일한 시간정보로 사용한 연구들[2-7]과는 달리 기사에서 나타난 절대시간 및 상대시간을 자동 추출한 후 이를 기사의 발행일을 기준으로 정규화 하고, 기사에 보도되는 사건을 대표하는 자질중 하나로 사용하였다.

3. 시간정보 추출 시스템

시간정보 추출 시스템은 기사가 보도하는 사건과 관련된 시간을 보다 정확하게 추출하기 위해서, 기사 내 시간표현을 자동으로 인식하고 특정 시각이나 기간으로 정규화하는 방법[17]을 사용한다. 그림 1은 시간정보 추출과정을 보여주고 있다. 기사 내 시간표현은 FSA에 정의된 품사패턴과 비교를 통해 빠르게 선택된 후, 어휘사전을 사용하여 시간을 표현하고 있는지 여부를 결정하게 된다. 이렇게 추출된 시간표현은 기사의 발행일을 기준으로 정규화 규칙과 어휘사전을 이용해서 'YYYYMMDD' 형태의 절대시간으로 변환된다. FSA와 어휘사전은 '학습문서'를 통해 수작업으로 구축된다.

기사는 '고유번호', '제목', '발행일', '내용', '작성자' 정보로 구성되어 있으며, 시간정보 추출시스템은 기사 내 절대시간 정보는 물론 기사의 '발행일'을 기준으로 상대시간 정보를 정규화한다. 시스템의 처리 결과로서, 기사 내 시간표현이 나타난 문장마다 시각 표현인 경우 'YYYYMMDD'를, 기간 표현인 경우 'YYYYMMDD-YYYYMMDD' 형태의 정규화된 시간표현을 표시하게 된다.

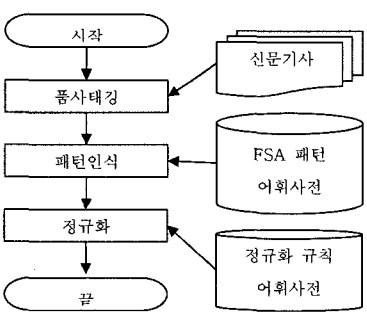

그림 1 시간정보 추출 단계

3.1 확장 태그를 이용한 품사태깅

시간표현에 사용된 패턴을 단순화하기 위해서 FSA는 시간표현에 사용되는 직접적인 단어가 아닌 단어의 품사로 구성되었다. 품사태그는 시각과 기간에 사용되는 어휘를 구분하기 위해서 시각을 의미하는 't(time)'와 기간을 의미하는 'd(duration)'를 기존 태그의 뒤에 표기하는 방법으로 확장되었다. 표 1은 패턴에 사용되는 태그의 의미와 해당 태그에 속하는 어휘중 일부를 보여주고 있다.

① 기사 내용에 대해서 품사태거를 사용해서 각 단어별 품사정보를 태깅한다.
② 시각표현 어구인지 판별한다.
  A. 미리 생성된 시각 FSA와 어구의 패턴 비교를 수행한다.
  B. 시간표현 가능 어구인 경우 어휘 사전에 등록된 각 품사별 어휘와 비교를 통해 시간표현 어구를 선별한다.
③ 기간표현 어구인지 판별한다.
  A. 미리 생성된 기간 FSA와 어구의 패턴 비교를 수행한다.
  B. 시간표현 가능 어구인 경우 어휘 사전에 등록된 각 품사별 어휘와 비교를 통해 시간표현 어구를 선별한다.
④ 정규화 수행
  A. 절대시간의 경우 정규화를 수행한다.
  B. 상대시간의 경우 기사의 발행일을 기준으로 정규화 규칙을 적용해서 절대시간으로 정규화를 수행한다.
  C. 특정 어휘사전에 등록된 단어들은 단어별 시간표현 정보를 기사의 발행일을 기준으로 절대시간으로 정규화를 수행한다.

그림 2 시간정보 추출 과정

표 1 태그별 의미와 어휘

태그	의미	어휘
SCD	기호나 숫자	1994, 11, …
NNBU_t	단위성 의존명사	년, 월, 일, …
NNCG_t	보통일반명사	새해, 올해, 어제, 오늘, …
NNCG_d		상반기, 하순, 월말, 봄,…
NPI_t	지시대명사	이날, 당일, …
NNP_t	고유명사	단오, 크리스마스, …
NNP_d		한일월드컵, …
PX	보조사	부터, 까지, …
XSNN	접미사	말, 초, …
DU	수관형사	한, 두, 첫, …
PA	부사격 조사	에, 에는, …

시간표현 추출과정에서는 시각표현을 추출하기 위해 4개의 FSA가 구축되었고, 기간표현을 추출하기 위해 5개의 FSA가 구축되었다. 그림 3은 시각을 추출하기 위해 사용되는 4개의 FSA를 보여준다. P1은 "1월에는"과 같이 숫자와 일반명사가 사용된 어구를 "1(SCD)" + "월(NNBU_t)" + "에는(PA)"으로 인식하기 위해 사용된다. P2는 "2000년 1월에는"과 같은 어구를 "2000(SCD)" + "년(NNBU_t)" + "1(SCD)" + "월(NNBU_t)" + "에는(PA)"와 같이 처리하기 위해 사용되며, P3은 '단오'와 고유명사가 조사와 같이 사용된 경우 이를 인식하기 위해 사용된다. P4는 "첫 달"과 같이 관형사와 단위명사가 같이 사용된 어구를 "첫(DU)" + "달(NNBU)"로 인식하기 위해 사용된다. 그림 4는 기간을 추출하기 위해 사용되는 5개의 FSA를 보여준다. D1은 "1월부터"와 같이 시각을 나타내는 정보와 기간 보조사가 같이 사용된 어구를 "1

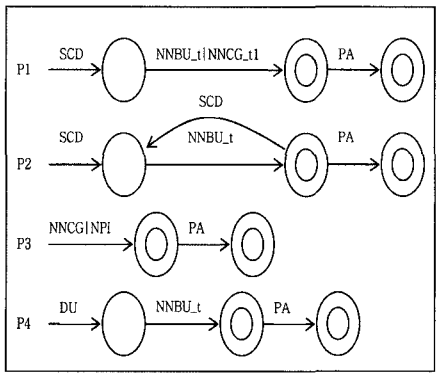

그림 3 4개의 시각 FSA

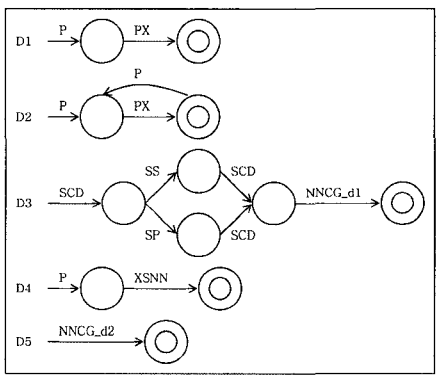

그림 4 5개의 기간 FSA

월(P1)" + "부터(PX)"로 인식하기 위해 사용된다. D2는 "1월부터 3월까지"와 같이 D1의 패턴이 반복되는 경우 "1월부터(D1)" + "3월까지(D1)"로 처리하기 위해 사용된다. D3는 "1.4 분기"와 같이 단어나 숫자가 기간을 나타내는 일반명사와 같이 사용된 경우 "1(SCD)" + ".(SS)" + "4(SCD)" + "분기(NNCG_d)"로 인식하기 위해 사용된다. D4는 "1월초"와 같이 시각을 나타내는 정보와 기간을 나타내는 접미사와 같이 사용된 경우 "1월(P1)" + "초(XSNN)"로 인식하기 위해 사용되고, D5는 "상반기"와 같이 일반명사가 단독으로 사용된 경우에 "상반기(NNCG_d)"로 인식하기 위해 사용된다.

3.2 어휘사전

시간표현 어구 추출시 FSA만 사용하는 경우, 품사패턴은 일치하지만 시간을 표현하고 있지 않는 어구들이 추출될 수 있다. 예를 들면 P3의 경우 "자동차(NNCG)+에(PA)"와 같은 패턴이 FSA만 사용하는 경우 시간표현 어구로 등록될 수 있다. 특히 "식목일", "단오" 등의 단어들과 같이 보통 일반명사나 고유명사가 특정 시각이나 기간을 나타내는 경우 시간표현에 사용되지 않는 다른 일반명사나 고유명사들과 구분이 요구된다. 이런 문제를 해결하기 위해서 시간표현에 사용되는 어휘를 사전으로 구축하였다. 어휘사전은 시간표현 어구 추출시 FSA를 보완하고, 추출된 시간표현을 정규화하는 과정에서 정규화 규칙들과 같이 사용된다. 어휘사전은 일반 어휘사전과 특정 어휘사전으로 구분되며, 일반 어휘사전은 시간표현에 사용되는 일반적인 단어들이 품사와 의미별로 분류되어 있고 특정 어휘사전은 특정일이나 특정기간을 표현하는 단어나 구로 구성되어 있다. 표 2와 표 3은 일반 어휘사전에 속하는 단어들을 품사와 의미별로 분류하여 보여주고 있다.

표 2 시간어휘 범주

범주		시간어휘
관형사		이, 한, 두, …
숫자		1994, 12, …
시간명사	년	년, 올해, 작년, 후년,…
	월	월, 섣달, 정월, 지난달, …
	일	일, 오늘, 어제, 모레, …
조사		부터, 까지, …
접미사		초, 말, 간, …

표 3 기간 표시어 분류

범주	기간 표시어
시작 추측	부터, 이래, 이후
종결 추측	까지, 안에
범위 추측	초, 동안, 상반기, 분기, …

특정 어휘사전에 속하는 각각의 단어나 구는 특정일 또는 특정기간으로 정규화할 수 있으며, 어휘별로 "반복여부(0|1):음력(0|1):년월일(YYYYMMDD):주(0~5):일(0~7)"의 형태를 사용해서 시간을 표현한다. 이 정보를 사용해서 어휘가 가리키는 특정일 또는 특정기간이 매년 반복되는지, 음력 또는 양력인지, 언제인지, 몇째 주인지, 며칠인지를 파악할 수 있다. 따라서 특정 어휘사전은 기사에 나타난 특정 어휘를 대상으로 특정 날짜와 매핑하는 작업을 통해 시간표현을 추출하고 정규화하는데 도움을 준다. 특정 어휘사전은 한국의 명절을 가리키는 단어는 물론 특정 날짜나 기간을 나타내는 고유명사 등을 포함하고 있으며, 총 117개의 어휘가 등록되어 있다.

표 4는 특정 어휘사전에 등록된 단어와 시간을 표현에 사용되는 정보구조의 예를 보여준다. "식목일"에 대한 시간표현을 살펴보면, "식목일"은 매년 반복되며(1), 양력을 기준으로 하고(1), 연도 없이 월과 일만 지정되어 있으며(00000405), 몇째 주인지 정해져 있지 않고(0) 또한 며칠(0)인지 정해져 있지 않기 때문에 "1:1:00000405:0:0"로 표현되는 것을 알 수 있다. "근로자의 날"에 대한 시간표현을 살펴보면, "근로자의 날"은 매년

반복되며(1), 양력을 기준으로 하고(1), 연도 없이 월만 지정되어 있으며(00000500), 매주 첫째주(1) 월요일(1)로 정해져 있기 때문에 "1:1:0000:5:0:1:1"로 표현되었다.

표 4 특정 어휘사전

단어	시간 표현
식목일	1:1:00000405:0:0
4.19혁명기념일	1:1:00000419:0:0
단오	1:0:0000:5:5:0:0
하지	1:1:0000:6:22:0:0
한일월드컵	0:1:20020531:0:0 ~ 0:1:20020630:0:0
근로자의 날	1:1:0000:5:0:1:1

3.3 정규화 규칙

기사 내에서 자동 추출된 시간은 기사의 발행일을 현재일로 지정한 후 표 5와 같은 어휘별 정규화 규칙을 사용해서 절대시간(YYYYMMDD)으로 변환된다. 정규화 과정에서는 시간표현 어구를 탐색하면서 시간정보가 나오면 절대시간으로 변환하는데 이때 나타나지 않는 시간단위는 '0'으로 표기된다. 시간표현 어휘들이 연속되어 나타나는 경우 앞 단어보다 뒤에 나온 단어의 시간 단위가 작은 경우 앞 단어에서 기록된 절대시간을 수정해서 뒷 단어가 나타내는 시간을 절대시간에 반영한다. 추출된 시간표현 어구는 시각을 나타내는 경우 '시각'으로 정규화하고, 기간을 나타내는 경우 '시작시각 : 종료시각'으로 정규화 된다. 즉 2000년 3월 20일 발행된 기사에서 '올해 3월 15일'이라는 어구가 시간표현 어구로 추출되면 '올해'라는 단어가 절대시간 '20000000'로 변환된다. 이후 연속된 시간표현 단어 '3월'은 앞서 나타난 '올해'보다 시간단위가 작기 때문에 절대시간 '20000300'로 수정된다. 마찬가지로 연속된 시간표현 단어인 '15일'은 앞서 나타난 '3월'보다 시간단위가 작기 때문에 절대시간 '20000315'으로 수정된다. 정규화에서 표현하는 시간의 최소단위는 '일' 단위이기 때문에 절대시간이 '일'까지 확장되면 이후 절대시간을 수정하는 작업을 중지하고 이를 시각으로 사용한다. 특정일이나 특정기간과 관련된 어휘의 경우는 특정 어휘사전을 참조하여 절대시간으로 직접 변환하며, 이외의 시간정보는 정규화 규칙을 사용해서 변환하게 된다.

표 5 시간 정규화 규칙

시간	어휘	정규화 규칙
과거	작년, 지난해, 전년, …	현재년도 - 1
	지난달, …	현재월 - 1
	구랍, …	현재년도 - 1 년 12월
	어제, 전날, …	현재일 - 1
현재	올해, 금년, 올, …	현재년도
	이달, …	현재월
	이날, 오늘, 당일	현재일
미래	내년, …	현재년도 + 1
	후년, …	현재년도 + 2
	내달, 다음달, …	현재월 + 1
	내일, …	현재일 + 1
기간	월초	현재월 1일 ~ 현재월 10일
	상반기	현재년도 1월 ~ 현재년도 6월

4. 평가

시간정보 추출시스템의 성능을 평가하기 위해 제안된 방법을 사용해서 한국어 신문기사를 대상으로 시간정보를 추출하고 정규화하였다. 또한 기사에서 추출된 시간정보를 사용하는 경우 TDT 시스템의 성능에 미치는 영향을 알아보기 위해 사건을 지정하고 사건별 관련기사를 찾는 과정에서 기사에서 추출된 시간정보를 사용하는 실험을 하였다. 제안된 방법의 평가를 위해 다음과 같은 세가지 실험을 수행하였다.

- 시간추출 시스템 전체에 대한 성능 평가
- 시간추출 시스템 단계별 성능 평가
- TDT 시스템에서 자동 추출된 시간의 유용성 평가

FSA와 어휘사전, 정규화 규칙은 한국경제신문과 한국일보 신문기사[18]를 대상으로 구축되었고, 평가는 조선일보 신문기사를 대상으로 진행되었다. 정치, 경제, 사건/사고 등 다양한 범주로부터 25개의 사건이 선정되었고, 정보검색 시스템을 사용해서 각각의 사건 기술 문서로부터 키워드를 추출하여 검색 질의를 생성한 후 검색을 수행하였다. 사건별 검색 질의를 수행한 결과로 생성된 문서 집합을 대상으로 평가자가 사건 관련기사 여부를 판정하였다. 약 11만 건에 해당되는 기사 중에서 25개의 사건과 관련된 1,320건의 기사가 평가자에 의해 최종 선정되었고, 이들을 대상으로 시간정보 추출과정의 정확도를 평가하였다. 선정된 25개 사건은 사건 발생일 기준으로 최소 0일(예: "검찰 피의자 구타 사망사건(20021026)"과 "체첸 반군 모스크바 극장 인질사건(20021023-20021026)")에서 최대 55일(예: "은행권 주5일 근무제 실행(20020706)"과 "태풍 루사 한반도 강타(20020830)")의 시간적 간격을 가진다. 표 6은 25개의 사건에 대해 각각 사건 번호와 사건명 그리고 사건별 정답기사수를 보여준다. 사건별로 최소 7건에서 최대 204건의 기사가 사건 정답기사로 할당되었으며, 하나의 기사는 하나의 사건에 대해 보도한다는 가정하에 기사의 주요내용이 25개의 사건과 관련되었을 경우 해당 사건에 대한 정답기사로 선정하였다. 25개의 사건 중 일부는 동일한 날짜에 발생하였다.

4.1 시간추출 시스템 평가

시간추출 시스템에 대한 평가는 사람에 의해 추출된

표 6 사건 리스트

번호	사건명	정답기사수
1	군산시 유흥가 화재	22
2	미국 악의축 국가로 북한 선언	141
3	김동성-오노 동계올림픽 판정 시위	81
4	서울 상봉동 은행에 총기 강도 사건	17
...	...	...
25	손기정옹 타계	20
계		1320

결과와 제안된 방법을 사용해서 추출된 결과의 비교를 통해 이루어졌다. 평가자들은 1,320건의 기사에서 모든 시간표현을 추출하고 이를 절대시간으로 표현하였다. 그 결과 기사당 0~2 문장이 시간정보를 포함하고 있었고, 총 1,514개의 문장에 2,211개의 시간이 발견되었다. 시간추출 시스템에 의해 2,109의 시간정보가 추출되고, 그 중 1,847개가 정답과 일치하여 전체적으로 83.5%의 재현율과 87.5%의 정확도를 보여주었다.

시간추출 시스템의 단계별 성능을 알아보기 위해 시간표현 추출 단계와 정규화 단계로 구분하여 재현율과 정확도에 대한 실험을 수행하였다. 시간표현 추출 단계에서 FSA와 어휘사전을 이용한 시간표현 추출은 전체 2,109개중 1,981개가 추출되어 90%의 재현율과 94%의 정확도를 보여주었다. 오류의 유형을 빈도순으로 살펴보면 첫째, 잘못된 품사태깅으로 인해 발생하는 오류(70%), 둘째 어휘사전에 등록되지 않은 어휘로 인해 발생하는 오류(20%), 마지막으로 시간정보가 새로운 패턴으로 사용된 경우 발생하는 오류(10%)이다.

정규화 과정에서는 자동 추출된 1,981개의 시간표현중 1,847개가 정확하게 정규화되어 93%의 정확도를 보였으며, 대부분의 오류는 상대시간으로 표현된 어구를 절대시간을 변환하는 과정에서 발생하였다. 오류의 유형을 빈도순으로 살펴보면 다음과 같다. 첫째, 문서의 의미를 파악하지 않기 때문에 생기는 오류(70%)로서 기사의 발행일이 정규화 규칙을 적용하는 기준일로 사용되지 않은 경우이다. 예를 들면 2000년 3월 10일 발행된 기사에서 2000 3월 9일에 일어난 사건내용을 보도하는 경우 기사에 '사건 당일'이라는 시간표현 어구가 사건이 발생한 2000년 3월 9일을 기준으로 정규화되지 않고, 기사가 발행된 2000년 3월 10일을 기준으로 '당일' 이라는 단어가 2000년 3월 10일로 정규화되면서 발생하는 오류이다. 둘째, 시간표현 어구에 대한 정규화 규칙이 없어서 정규화를 수행하지 못한 경우(30%)이다.

4.2 추출된 시간정보의 유용성 평가

TDT 연구는 크게 사건 탐지를 위한 과정(event detection)과 사건 추적을 위한 과정(event tracking)으로 구분할 수 있으며 사건 탐지를 위한 과정은 문서 집합을 사건별로 분류하는 작업이고, 사건 추적을 위한 과정은 주어진 기사를 사용해서 관련기사를 찾는 작업이다. 사건이 발생한 시간과 공간이 다른 경우 서로 다른 사건으로 구분할 수 있기 때문에, 사건과 관련된 시간정보는 유사한 사건들에 대한 보도 기사를 구분하기 위한 단서로 사용되고 있다. 대부분의 TDT 연구에서는 기사를 발행일순으로 정렬한 후 기사간 인접도를 유사도에 반영하는 방법으로 시간정보를 활용하고 있다[2-6].

기사의 발행일과 사건 발생일은 다를 수 있으며, 또한 사건 관련기사가 시간의 차이를 두고 보도되는 경우 사건일과 기사의 발행일은 달라지게 된다. 기사들은 이전 사건과의 연관성을 나타내기 위해 기사 내용에 사건일 또는 지명등을 사용한다. 예를 들면 2000년 3월 20일 기사에서 같은 해 1월에 발생된 사건에 대해 보도하는 기사의 내용을 살펴보면 "지난 1월 군산에서 발생한 화재의 피해자가" 등의 어구를 사용해서 기존 사건과의 연관성을 나타내고 있음을 알 수 있다. 따라서 기사에서 보도되는 사건과 관련된 시간정보를 추출하고 이를 활용한다면 TDT의 성능을 향상시킬 수 있을 것이다. 특히 사건 추적을 위한 과정에서는 추적 단서로 주어지는 문서의 수가 매우 적기 때문에 기사에서 특정 시간이 나타난 경우 이를 단서로 활용해서 관련기사를 찾는다면 보다 정확한 결과를 얻을 수 있을 것이다.

실험에 사용된 1,320건의 문서에서 추출된 시간을 분석해보면 약 90%(1,181건)의 기사에서 기사의 발행일과 다른 시간정보가 추출되었으며, 이중 사건과 관련된 시간정보가 추출된 기사는 76%(1,009건)에 이르는 것을 알 수 있었다. 즉 전체 기사의 76%가 기사의 발행일과는 다른 사건의 발생일과 관련된 시간정보를 포함하고 있음을 알 수 있다. 따라서 기사의 발행일을 사용하는 경우와 기사에서 추출된 시간을 사용하는 경우 TDT 시스템의 성능에 영향을 미칠 수 있음을 알 수 있다. 본 실험에서는 기사에서 2개 이상의 시간정보가 추출된 경우 하나라도 기사의 발행일과 일치하지 않으면 기사의 발행일과 다른 시간정보를 가진 기사로 분류하였으며, 추출된 시간정보가 사건 발생일자 전후 1주일 또는 사건 발생일자를 포함하는 경우 사건과 관련된 시간정보

표 7 발행일별 사건 관련기사의 분포

사건번호	관련기사	1주	2주	3주	그외
1	22	15	4	1	3
2	141	73	22	3	46
3	81	79	2	0	0
4	17	9	0	8	0
...		...	...		...
25	20	20	0	0	0
계	1295	786 (0.61)	184 (0.14)	64 (0.05)	261 (0.20)

가 추출된 기사로 분류하였다. 따라서 기사에 나타난 시간정보가 기사와 사건의 연계성을 파악하는데 중요한 단서가 되는 것을 알 수 있다.

사건은 크게 "예측 가능한 사건"과 "예측이 불가능한 사건"으로 구분할 수 있으며, 실험에 사용된 총 25개의 사건중 "은행 주 5일 근무제 시행"과 같이 미리 지정된 시각에 발생하는 사건은 2건이 포함되어 있다. 총 1,320건의 사건 관련기사중 25건이 사건 발생일전에 보도되어 정답 기사로 판정되었지만, 사건 정답기사는 사건 발생일 이후에 나타날 수 있다는 가정아래 사건일 이후의 1,295건의 기사를 사건 정답기사로 선정하고 이후의 실험을 진행하였다. 표 7은 사건 정답기사를 사건 발생후 1주일, 2주일, 3주일, 그외로 구분하여 해당 기간에 발행된 정답기사의 수를 보여주고 있다. 총 1,295건의 정답기사는 사건 발생일을 기준으로 1주일안에 61%(786건), 2주일안에 75%(970건), 3주일안에 80%(1,034건)로 분포되어 있음을 알 수 있었다. 사건 발생후 3주안에 전체 관련기사의 80%가 나타나므로 기존의 TDT 연구에서 정해진 기간에 해당되는 기사들을 동일 클러스터에 할당하기 위한 노력은 타당하다고 볼 수 있다. 그러나 관련기사의 20%가 사건 발생후 3주일 범위에 속하지 않기 때문에 단순히 기사의 발행일만 고려해서 시간 윈도우를 정해고, 정해진 시간 윈도우에 속한 기사들만을 대상으로 정답기사를 모두 찾기는 쉽지 않다. 또한 기사가 시간 차이를 두고 전개되는 경우 모든 정답기사를 포함하기 위해 시간 윈도우의 범위를 확대할수록 시간에 따른 사건간 변별력이 떨어지게 되므로 정답기사 판정에 오류가 많아지면서 정확율은 떨어지게 된다. 따라서 기사에서 사건과 관련된 시간을 추출한 후 사건일과 비교를 통해 동일 사건에 대한 기사 여부를 결정하는 것이 필요하다.

4.3 추출된 시간정보의 유용성

기사에서 추출된 시간을 사용하는 경우 TDT 시스템의 성능에 미치는 영향을 판단하기 위해 다음과 같이 세가지 경우를 설정하여 실험을 진행하였다. 25개의 사건에 대한 정답기사 1,295건을 약 11만건의 기사 집합에서 찾는 과정에서 각각의 경우에 대한 정확도와 재현율을 측정하였다.

- 기사 내용만 사용한 경우 (A)
  수작업을 통해 각각의 사건에 대해 중심값을 지정한 후 약 11만건에 해당되는 전체기사를 분류하였다. 사건 중심값과 기사의 유사도가 지정된 임계값 이상일 경우 기사는 사건 클러스터에 중복이 가능하게 할당되며, 임계값은 각 사건별로 모든 정답 기사를 포함할 수 있도록 조정되었다. 모든 정답기사를 찾는 과정에서 발생하는 오류문서의 수를 측정하였다.
- 기사의 내용과 발행일을 사용한 경우 (B)
  (A)를 통해 생성된 정답기사 집합을 기사의 발행일 순으로 정렬한 후, 사건 발생일부터 기사의 발행일이 가장 먼 정답기사를 포함할 때까지 시간 윈도우를 단계별로 확장하면서 시간 윈도우내에서 정답을 찾는 방법으로, 지정된 시간 윈도우에 포함되는 오류문서의 수를 측정하였다.
- 기사의 내용, 발행일, 기사에서 추출된 시간을 사용한 경우 (C)
  (B)와 동일한 방법으로 시간 윈도우를 확장하면서 지정된 시간 윈도우에 기사의 발행일 또는 기사에서 추출된 시간정보가 포함되는 경우를 사건에 대한 정답기사로 판정하였다. 기사에서 추출된 시간은 사건 발생일 또는 사건 발생후 1주일안에 해당되는 경우만 정답기사로 판정하였다.

표 8은 정답기사를 찾는 과정에서 각각의 사건에 대해 정답기사로 판정되는 기사의 수를 보여주고 있다. 총 1,295건의 정답기사를 찾는 과정에서 기사의 내용만 사용한 경우(A)는 5,067건이 정답기사로 판정(오류 문서의 비율이 74%), 기사의 내용과 기사의 발행일을 고려한 경우(B)는 2,513건을 정답기사로 판정(오류 문서의 비율이 48%), 기사의 내용과 발행일, 기사에서 추출된 시간을 고려한 경우(C)는 2,124건을 정답기사로 판정(오류 문서의 비율이 39%)하였다. 기사에서 자동 추출된 시간정보를 사용하는 경우가 기사의 발행일만 사용하는 경우보다 약 17%의 성능 향상을 가져왔다. 표 9는 모든 정답기사를 찾는 과정에서 사건 발생일을 기준으로 시간 윈도우를 사건 발생후 1주, 2주, 3주, 그 외로 확장하는 경우 기사의 내용과 발행일을 사용하는 경우(B)와 기사의 내용과 발행일, 추출된 시간정보를 사용하는 경우(C) 시간대별로 정답기사로 판정되는 기사의 수를 보여주고 있다. 표 10은 (B)와 (C)경우 시간대별로 정확도와 재현율을 측정한 결과를 보여주고 있다. 사건 발생후 1주만 살펴보면 기사에서 추출된 시간중 하나라도 이 범위에 속하면 정답기사로 처리되기 때문에 (B)경우에 비해 정확도는 약 5%로 떨어지지만, 기사의 발행

표 8 추출된 시간정보를 사용하여 얻는 성능 향상

사건번호	정답기사수	(A)	(B)	(C)
1	22	130	38	25
2	141	277	188	171
3	17	137	35	32
4	14	53	17	17
...	...	...	...	...
25	20	97	39	25
계	1295	5067	2513	2124
정확도		0.26	0.52	0.61(+17%)

표 9 시간대별 사건 관련기사의 분포

	1 주		2 주		3 주		그 외	
	전체	평균	전체	평균	전체	평균	전체	평균
정답 기사수	786	31	970	39	1034	41	1295	52
(B)	801	32	1305	52	1990	80	2513	101
(C)	1201	48	1422	57	1657	66	2124	85

표 10 시간대별 성능향상

	1 주		2 주		3 주		그 외	
	정확도	재현율	정확도	재현율	정확도	재현율	정확도	재현율
정답 기사	1.00	0.61	1.00	0.75	1.00	0.80	1.00	1.00
(B)	**0.98**	**0.61**	**0.74**	**0.75**	**0.52**	**0.80**	**0.52**	**1.00**
(C)	**0.93 (-5%)**	**0.86 (41%)**	**0.84 (14%)**	**0.92 (23%)**	**0.73 (40%)**	**0.93 (16%)**	**0.61 (17%)**	**1.00 (0%)**

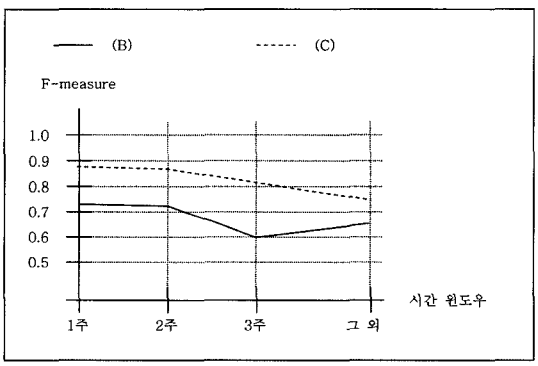

그림 5 F-measure 비교

일이 사건 발생후 1주에 속하지 않은 문서들이 기사에서 추출된 시간 때문에 이 범위에서 정답기사로 판단되기 때문에 재현율은 약 41% 증가하는 것을 알 수 있다.

그림 5는 기사의 내용과 발행일을 사용하는 경우(B)와 기사의 내용과 발행일, 추출된 시간정보를 사용하는 경우(C)에 대해 시간대별 F-measure 수치를 보여준다.

실험결과 일정한 시간 윈도우에 속하는 기사를 대상으로 정답기사를 찾는 과정에서 기사의 발행일만 사용하는 경우에 비해 기사에서 추출된 시간정보를 사용하는 경우 시간 윈도우의 확대에 따라 발생하는 오류 기사의 수를 줄일 수 있었다. 따라서 기사에서 자동 추출된 시간을 사용하면 필요 이상으로 시간 윈도우가 확대되는 것을 막아주기 때문에 TDT 시스템의 성능을 향상시킬 수 있다.

5. 결 론

본 논문에서는 TDT 시스템의 성능을 향상시키기 위해 기사에서 자동 추출된 시간정보를 사용하는 방법을 제시하였다. 그리고 기사의 발행일을 사용하는 경우와 기사의 발행일과 기사에서 추출된 시간을 사용하는 경우를 비교하여 실험하므로써 자동 추출된 시간정보의 유용성을 증명하였다.

제안된 시간표현 추출 방법은 단어의 품사정보로 구축된 4개의 시각 패턴과 5개의 기간 패턴을 어휘사전과 같이 사용해서 94%의 정확도와 93%의 재현율을 나타내었고, 이 과정에서 발생한 오류의 대부분은 품사태깅 과정에서 발생하였다. 추출된 시간표현은 어휘별 정규화 규칙과 어휘사전에 등록된 시간정보를 사용해서 절대시간으로 변환되는데, 이 과정의 정확도는 93%이고 발생한 오류의 대부분은 상대시간으로 표현된 시간정보를 정규화하는 과정에서 발생하였다. 이런 오류는 시간정보가 나타난 문장의 이해를 통해서 해결할 수 있다.

신문기사의 경우 기사의 특성상 기사의 발행일과 다른 시간을 나타내는 문장을 가진 기사가 전체의 90% 이상을 차지하고 있기 때문에, 기사가 보도하는 사건 발생일을 고려하지 않고 기사의 발행일만 고려하는 경우 그 정확성이 낮아질 수 밖에 없다. 또한 기사는 사건의 전개에 따라 시간 차이를 두고 보도되는 경우가 많기 때문에, 동일 사건과 관련된 모든 기사가 연속되어 나타나지 않을 수 있다. 따라서 기사를 대상으로 동일 사건에 대한 보도기사인지 아닌지 구분하는데 시간은 매우 중요한 역할을 하며, 기사별로 정확한 사건 발생일을 추출하여 사용한다면 TDT 시스템의 성능을 향상 시킬 수 있을 것이다. 추출된 시간의 유용성을 증명하기 위해 사건 정답기사를 찾는 과정에서 기사의 내용만 사용해서 정답기사를 찾는 경우와 기사의 발행일을 사용하는 경우, 기사에서 추출된 시간을 사용하는 경우로 구분하여 실험을 하였다. 실험에서 사건 발생후 3주까지의 시간 윈도우내에서 나타난 결과를 살펴보면 기사에서 자동 추출된 시간정보를 사용하는 경우가 기사의 발행일만 사용하는 경우보다 정확도는 40%, 재현율은 16% 향상됨을 알 수 있었다.

기사에서 추출된 시간을 사용하는 경우 사건과 관련

되지 않은 시간정보가 추출되어 발생하는 오류를 줄이기 위한 방법으로 추출된 시간이 사건과 관련이 있는지를 판단할 수 있는 방법이 요구된다. 즉 기사에서 추출된 시간이 사건과 관련있는지를 판단하거나 또는 다수의 시간이 추출된 경우 이 중에서 사건과 관련된 시간을 선택하는 방법은 시스템의 정확도는 향상시킬 수 있을 것이다. 따라서 향후 연구로 사건과 동사를 연결한 후 시간정보를 동사와 연결하는 방법[17]과 같이 사건과 관련된 정확한 시간정보를 선택하는 연구가 진행되어야 하겠다.

참 고 문 헌

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In Proceedings of the ACM SIGIR conference on Research and development in information retrieval, Pages 10-18, 2001.

[2] J. Allan, R. Papka and V. Lavrenko. On-line new event detection and tracking. In Proceedings of ACM SIGIR conference on Research and development in information retrieval, Pages 37 - 45, 1998.

[3] J. Allan et al. Topic Detection and Tracking Pilot Study Final Report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Feb 1998.

[4] Y. Yang, T. Pierce, and J. Carbonell. A Study on Retrospective and On-Line Event Detection. In Proceedings of ACM SIGIR Conference on Research and development in information retrieval, Pages 28-36, 1998.

[5] Y. Yang et al. Learning Approaches for Detection and Tracking News Events. IEEE Intelligent Systems, 14(4):32-43, July/August 1999. Special Issue on Applications of Intelligent Information Retrieval.

[6] R. Papka, J. Allan, and V. Lavrenko. UMASS approaches to detection and tracking at TDT2. In Proceedings of the TDT-99 workshop. NIST, 1999.

[7] T. Leek, R. Schwartz and S. Sista. Probabilistic Approaches to Topic Detection and Tracking. TOPIC DETECTION AND TRACKING, Kluwer Academic Publishers, Pages 67-83.

[8] B. Sundheim, N. Chinchor, Named Entity Task Definition, Version 2.0, 31 May 95. In Proceedings of the 6th Message Understanding Conference (MUC-6). Pages 319-332, Morgan Kaufman Publishers, Inc., 1995.

[9] N. Chinchor. MUC-7 Information Extraction Task Definition, Version 5.1, 23 July 1998. In Proceedings of the 7th Message Understanding Conference (MUC-7), 1998.

[10] L. Ferro, I. Mani, B. Sundheim, G. Wilson. TIDES Temporal Annotation Guidelines. MITRE Technical Report Version 1.0.2, June 2001.

[11] G.B. Alice, T. Meulen. Representing Time in Natural Language. MIT Press, Cambridge, Massachusetts, 1995.

[12] B. Moulin. Temporal Contexts for Discourse Representation: An Extension of the Conceptual Graph Approach. Artificial Intelligence, 7: Pages 227-255, 1997.

[13] Juntae Yoon, Yoonkwan Kim, Mansuk Song. Identifying Temporal Expression and its Syntactic Role Using FST and Lexical Data from Corpus. In Proceedings of Colling, 2000.

[14] J. Allan. Introduction to Topic Detection and Tracking. TOPIC DETECTION AND TRACKING, Kluwer Academic Publishers, Pages 1-16.

[15] N. Stokes, P. Hatch, J. Carthy. Lexical semantic relatedness and online new event detection. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Pages 324-325, 2000.

[16] F. Fukumoto, Y. Suzuki. Event Tracking based on Domain Dependency. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Pages 57-64, 2000.

[17] Pyung Kim, Kiyoun Sung, Sung Hyon Myaeng, Jae Cheol Ryou. Extracting Temporal Information from Korean News Articles for Event Detection and Tracking. In Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Pages 392-401, 2003.

[18] Lee, S. H., Myaeng, S. H. Kim, J. Y., Jang, D. H., Seo, J.H., Kim, H. Packaging Hanguel Test Collection as an Evaluation System of Information Retrieval. In Proceedings of the 5th Korea Science & Technology Infrastructure Workshop (in Korean). 2000.

김 평

1997년 충남대학교 전산학과(학사). 1999년 충남대학교 컴퓨터과학과(석사). 2004년 충남대학교 컴퓨터과학과(박사). 2004년~현재 한국과학기술정보연구원 NTIS 사업단 선임연구원. 관심분야는 정보검색, 자연어 처리, 시멘틱 웹

맹 성 현

1983년 미국 캘리포니아 주립대학 학사 1987년 미국 Southern Methodist University(SMU) 석사 및 박사. 미국 Temple University 조교수. Syracuse University 종신교수. 충남대학교 교수 역임. 현재 한국정보통신대학교(ICU) 공학부 교수 관심분야는 정보검색, 텍스트마이닝, 디지털도서관, 시맨틱 웹 등임. 2002년 ACM SIGIR Conference Program Committee Chair, AIRS 2004 Program Committee Chair, Information Processing & Management, Journal of Natural Language Processing, Journal of Computer Processing of Oriental Languages 편집위원 등으로 활동. Home page: http://ir.cnu.ac.kr.