사건 탐지/추적을 위한 시간 정보 추출

충남대학교 컴퓨터 과학과,[1] 주식회사 뱅크타운,[2] 한국정보통신대학원대학교[3]
김 평[1] · 성기윤[2] · 맹성현[3]

Temporal Information Extraction from Korean News for Event Detection and Tracking

Pyung Kim,[1] Ki Youn Sung,[2] Sung Hyon Myaeng[3]
Department of Computer Science,[1] Chungnam National University, Daejeon, Korea
Bank Town,[2] Inc., Korea
Information & Communications University,[3] Daejeon, Korea

요 약

시간정보는 사건 탐지/추적 시스템은 물론 정보 추출, 질의/응답 시스템 등에서 매우 중요한 역할을 한다. 본 연구에서는 한국어 신문 기사를 대상으로 시간 표현을 추출하고 정규화한 후 사건 관련 동사와 연결하는 자동화된 방법들을 제안하였다. 시간 표현을 추출하기 위해서 품사정보로 구축된 패턴과 시간 표현 어휘가 사용되었고, 정규화 과정과 사건 관련 동사와의 연결을 위한 규칙이 만들어졌다. 한국어 신문을 대상으로 제안한 방법의 단계별 평가를 수행하였고, 제안하는 방법의 확장성을 보이기 위해 서로 다른 도메인에도 실험을 하였다.

서 론(1장)

시간정보는 자연어 처리의 응용분야는 물론 정보검색의 응용에도 매우 중요하게 사용된다. 질의응답(QA : Question Answering) 시스템에서는 '언제'에 대한 답변으로, 정보추출(IE : Information Extraction) 시스템에서는 시간 정보 템플릿(template)을 채우기 위해서, 자동요약 시스템에서는 요약정보 생성에, 토픽 탐지추적(TDT : Topic Detection & Tracking) 시스템에서는 사건의 탐지 및 추적에 시간정보를 사용하고 있다. 한국어의 경우 이러한 응용과 관련되어 시간 정보를 자동 추출하고 이에 대해 정규화 하는 연구는 거의 수행되지 않았다.

외국의 경우 뉴스기사나 방송자료를 대상으로 사건을 지정하고 해당 사건에 대한 관련기사를 추적하거나 새로운 사건에 대한 기사를 자동탐지 하는 연구가 진행되고 있다.[1-4] 이 분야에서 사건은 "something that happens at particular time and place"으로 정의된다.[1] 즉 시간과 장소에 따라 동일한 사건인지 또는 서로 다른 사건인지를 구분할 수 있다. 정확한 시간 추출은 시간의 발생 순으로 기사를 정렬하여 시간별 사건추이를 이해할 수 있게 해 주는 것은 물론 사건의 추적과 탐지에도 중요한 역할을 한다.

대부분의 TDT 시스템에서는 특정 사건에 대한 기사는 일정기간에 보도되고 그 기간이 경과한 후의 기사는 새로운 사건에 대한 기사라고 가정한다. 이러한 가정에 입각한 사건의 특성을 반영하여 TDT 시스템에서는 문서(기사) 생성 날짜 순으로 문서집합을 정렬한 후 이 시간에 따른 차이를 가중치로 표현하여 같은 사건에 대한 기사인지를 판별하는 기준으로 사용하였다.[3,5,6]

본 연구에서는 뉴스 기사의 생성일을 기준으로, 기사에 언급된 실제 사건의 발생 시간정보를 추출하고, 그 결과를 정규화한 후, 사건 관련 동사와 연결을 함으로써 TDT 시스템의 성능 향상을 개선하는데 목표를 두었다.

2장에서는 사건정보의 추출을 위한 기존의 연구방법에 대해 기술하고, 3장에서는 본 연구에서 제시하는 사건정보의 추출 및 정규화 방법 그리고 추출된 시간과 사건동사와 연결하는 방법에 대해 기술한다. 4장에서는 신문 자료를 대

E-mail : pyung@cs.cnu.ac.kr
E-mail : yann@banktown.com
E-mail : myaeng@icu.ac.kr

상으로 수행한 실험에 대해 기술하고 5장의 연구 결과와 향후 연구방향에 대해 기술한다.

관련연구(2장)

정보추출(IE) 분야에서 문서로부터 구조 템플릿을 채우기 위한 하나의 방법으로 시간 정보를 추출하는 것과 관련된 연구가 진행되기 시작했다. MUC-6에서는 개체명 인식의 하부작업으로써 절대시간을 구분짓는 연구가 이루어졌고,[7] MUC-7에서는 이를 확장하여 상대 시간까지도 개체명에 포함시켰다.[8] 그렇지만 MUC에서는 시간 정보와 관련된 연구는 제한적이었으며 성능도 좋지 않았다.

미국 DARAP의 TIDES 프로젝트에서는 Temporal Guidelines을 통해 시간정보 표현을 ISO8601 형식으로부터 유도된 Gregorian calendar 방식에 기초를 두고 'YYMMDDhhmmss' 형식으로 시간을 표시하도록 하였다. 시간 표현을 위한 방법은 어느 특정 시점과 범위나 특정 시점을 기준으로 한 기간표현으로 나타내었다.[9]

자연어에서 사건을 인식하고 이를 시간적 추이에 따라 나열하기 위한 연구는 자연어 처리 관점뿐만 아니라 인공지능 및 추론 분야에서도 많이 시도되었다.[10-12] 형식담화(formal discourse) 기반이나[10,13] 말뭉치(corpus) 기반[14]에 의거하여 문서로부터 시간 정보를 추출하기 위한 방법들이 큰 주류를 이루고 있다.

1. 형식담화 기반 접근방법

이 접근방법은 자연어에 나타나는 표현 방식을 이해하고 의미 체계를 연구하는 방법으로, 특히 문맥에 나타난 시간 정보를 바탕으로 사건의 흐름을 추론할 수 있게 한다. 실제 우리가 사용하고 있는 담화에서는 시간 어구가 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.

문장에 나타난 시간 부사 어구는 사건의 시간적 관계를 나타내는 중요한 역할을 한다. 특히 동사의 시제와 상은 사건이 시간적으로 어떻게 연관이 있는지를 나타내는 중요한 단서가 된다. Ter Meulen은 서로 다른 문맥에서 시간정보가 어떻게 다르게 표현되는지와 그 의미적 역할을 파악하는데 초점을 두고, 이를 DAT(Dynamic aspect trees)라는 구조를 통해 표현하였다.[10] Moulin는 자연언어에 나타난 객체들간의 관계를 분석하여 개념그래프(CG : Concept Graph)를 구축하고 이를 사용하여 시간정보를 추출하는 연구를 수행했다.[13] 그러나 DAT나 CG 방법은 모두 자연어 문장의 의미를 표현하기 위해 드는 비용이 너무 높다는데 제약이 있다. 또한 다양한 언어적 표현집합, 즉 사전을 구축하고 이를 특성에 따라 범주화 할 수 없다는데 제약을 가진다.

2. 말뭉치 기반 접근방법

말뭉치로부터 구축된 어휘정보를 이용하여 문서내의 시간정보를 찾아내고, 그 시간정보의 문법적 역할을 활용하여 구문 분석시 발생하는 모호성을 제거하고자 하는 방법이 연구되었다.[14] 이 연구에서는 용어 색인기를 이용하여 시간 표현에 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 시간 표현은 FSA(Finite State Automata)를 통해 유효한 시간 표현으로 인식되고, 최종적으로 문장 내에서 해당 시간 표현의 문법적인 역할이 무엇인지를 알아낼 수 있다.

기존 연구들에서는 정보추출을 위해 대상 도메인을 한정한 상태에서 패턴 형태의 도메인 의존적인 정보를 구축하고 이를 이용해 텍스트의 특정 부분을 추출하는 방식을 주로 사용하고 있다.[7,8] 이를 다시 크게 두가지 방식으로 나누어 볼 수 있다. 첫째는 텍스트에서 개체명들을 인식하고 템플릿 엘리먼트, 템플릿 릴레이션, 시나리오 템플릿을 점차로 구성하면서 추출하고자 하는 정보를 획득하는 방식이고,[8] 둘째는 우선 텍스트에서 중요 부분을 추출한 후에 이를 대상으로 수동으로 만들어진 패턴과의 비교를 수행해 원하는 정보를 찾아내는 방식이다.[14] 첫번째 방식의 경우 각 단계에서 이용하는 도메인 정보 구축을 위해 우선 해당 도메인에서 중요시되는 정보들을 찾아야 하는 문제가 있다. 두번째 방식의 경우는 중요부분 추출 문제를 단순히 어휘정보에만 의존해 해결하고 있어 실질적인 정보 추출대상들을 효과적으로 추출해내지 못하고 있다는 점이다.

시간정보 추출 시스템(3장)

뉴스 기사나 방송자료를 대상으로 연구되는 TDT 시스템의 경우 기사나 방송자료의 발생일을 그 기사에서 언급된 사건에 대한 시간정보로 사용하고 있다.[3,5,6] 뉴스 기사나 방송 자료의 특성상 지난 사건에 대한 보도자료이기 때문에 기사의 발생일과 사건의 발생일에는 차이가 있을 수 있으며, 하나의 사건이 시간적 차이를 두고 기사에서 다뤄지는 경우도 발생할 수 있다. 이로 인해 시간과 공간정보로

구분되는 사건을 탐지하고 추적하는 TDT 시스템의 정확성이 낮아질 수 있다. 본 논문에서는 사건의 발생 시간정보를 보다 정확하게 추출하기 위해 기사 내부에 표현된 시간정보를 인식하여 사용하는 방법을 제시한다. 먼저 텍스트에 존재하는 시간정보는 구축된 어휘사전을 기반으로 FSA를 동작시켜 절대 시간 정보를 추출한다. 이렇게 추출된 시간정보는 사건을 묘사하는 동사와 연결된다.

시간표현 패턴, 어휘사전, 정규화 규칙과 사건 관련 동사 연결 규칙은 데이터 분석을 통해 생성되며, 이렇게 생성된 정보는 한국어 신문 기사를 대상으로 시간정보를 추출하는데 사용된다.

1. 시간 표현 패턴

시간 표현 패턴은 시간 표현에 사용되는 어구의 품사별 패턴을 사용하여 구축되었다. 직접적인 단어를 사용하지 않고 단어의 품사를 사용하여 패턴을 구성함으로써 일반화된 패턴을 구축할 수 있었다. 시간 표현 패턴은 시각을 표현하기 위한 패턴과 기간을 표현하기 위한 패턴으로 구분된다. 시각 표현 어구의 품사별 패턴은 유형에 따라 4개의 FSA로 구축되었고, 기간 표현 어구의 품사별 패턴은 5개의 FSA로 구축되었다.

Table 1은 패턴에 사용되는 품사 태그의 의미를 알려준다.

Fig. 1은 시각 정보를 추출하기 위해 구축된 4가지 패턴의 FSA를 보여준다. P1은 숫자와 단위보통명사가 같이 사용되어 "1월에는"과 같은 패턴을 위한 것이고, P2는 "93년 4월" 처럼 P1이 반복되어 사용된 경우를 나타낸다. P3는 하나의 단어로 특정시간을 나타내는 "단오"와 같은 단어를 위한 것이며, P4는 관형사와 의존단위명사가 같이 사용된 "이 달"과 같은 단어에서 시간정보를 추출하는데 사용된다.

Fig. 2는 기간정보 추출을 위해 구축된 5가지 FSA를 보여준다. D1은 시간을 나타내는 정보와 기간을 나타내는 보조사가 결합하여 "한 달 동안"과 같은 패턴을, D2는 시간을 나타내는 정보가 기간을 나타내는 보조사가 연속되어 사용된 경우로 "어제부터 오늘까지" 등의 형태를 인식한다. D3는 "1.4분기" 등의 형태를 인식하는 것으로 단어와 기호와 일반보통명사가 기간을 표현하는 경우이며, D4는 시간정보와 임의의 기간을 나타내는 접미사가 사용되어 "이달 초" 등의 형태를 추출하는데 사용된다. D5는 기간을 나타내는 일반명사가 단독으로 사용된 경우로 "상반기" 등의 형태가 있다.

Table 1. 품사설명

태그	의미	태그	의미
SCD	기호나 숫자	PX	보조사
NNBU	단위성 의존명사	XSNN	접미사
NNCG	보통일반명사	DU	수관형사
NPI	지시대명사	PA	부사격 조사
NNP	고유명사		

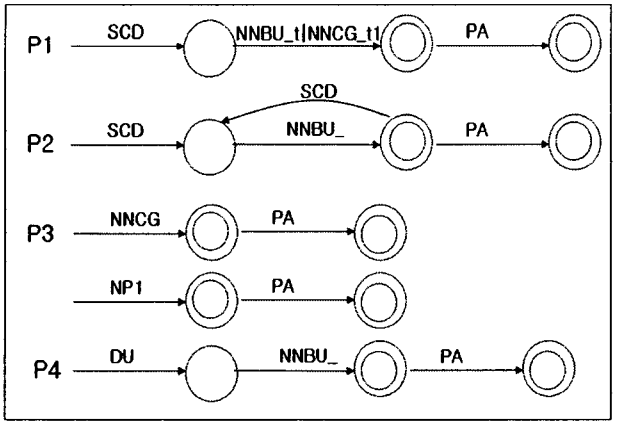

Fig. 1. 시각 FSA.

2. 어휘사전

품사 패턴으로 구성된 FSA 만으로 시간 표현어구를 추

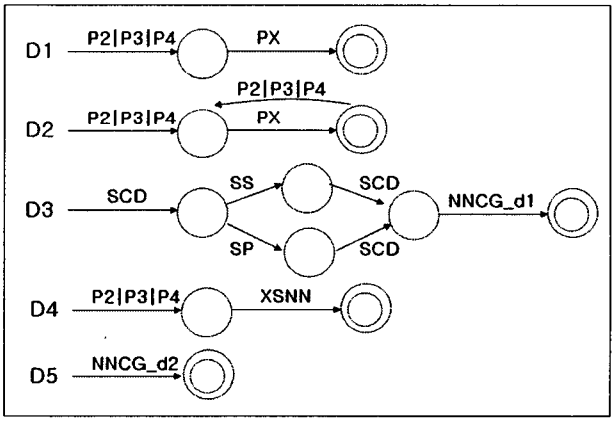

Fig. 2. 기간 FSA.

Table 2. 품사별 어휘

태그	어 휘
SCD	1994, 11, -, /, …
NNBU	년, 월, 일, …
NNCG	새해, 올해, …
NPI	이날, …
NNP	단오, 크리스마스, …
PX	부터, 까지, …
XSNN	말, 초, …
DU	한, 두, 첫, …
PA	에, 에는, …

출할 수 없기 때문에 어휘사전을 구축하였다.

어휘사전은 사용 목적에 따라 2가지로 구분되며, 하나는 시각과 기간 표현에 사용되는 품사의 어휘 별 사전이고, 다른 하나는 단어나 어구가 특정 시간이나 기간을 나타내는 경우로 단어나 어구와 이에 관련된 시간 정보가 같이 표현된 어휘 사전이다.

Table 2는 품사 태그별 의미와 시간정보를 표현하기 위해 사용된 어휘 사전을 보여준다.

Table 3과 4는 시간과 기간을 나타내기 위해 사용된 어휘를 범주별로 보여준다. 범주별 단어의 의미 정보는 패턴 패턴 추출 과정과 정규화 과정에서 사용된다.

Fig. 3은 하나의 단어나 어구가 특정일이나 특정 기간을 지정하는 경우, 이를 처리하기 위해 생성된 어휘사전의 예를 보여준다. 어휘사전의 어휘별 표현 형태를 살펴보면 다음과 같다. '특정일 : [무반복(0)|반복(1) : 음력(0)|양력(1) : 년도(YYYY) : 월(MM) : 일(DD) : 주(0~5) : 일(0~7)]'의 발생정보를 가지고 실제 뉴스 기사나 방송자료에서 나타난 일자정보를 대상으로 특정일과 매핑하는 과정을 수행한다. 즉 이 어휘사전에 해당되는 단어들은 정규화 과정에서 정규화 규칙과 상관없이 특정일로 매핑된다. 즉 특정 어휘사전은 기사에 나타난 특정 어휘를 대상으로 특정 날짜와 매핑하는 작업을 통해 시간표현을 정규화하는데 도움을 준다. 구축된 특정 어휘는 한국의 명절을 포함하여 특정 날짜나 기간과 매핑될 수 있는 고유명사를 포함하고 있으며, 전체 117개로 단어나 구로 구성되어 있다.

Table 3. 시간 어휘 범주

범 주		시간어휘
관형사		이, 한, 두, …
숫 자		1994, 12, …
시간명사	년	년, 올해, 작년, 후년, …
	월	월, 섣달, 정월, 지난달, …
	일	일, 오늘, 어제, 모레, …
조 사		부터, 까지, …
접미사		초, 말, 간, …

Table 4. 기간 표시어 분류

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

```
식목일 : [1 : 1 : 0000 : 4 : 5 : 0 : 0]
4.19혁명기념일 : [1 : 1 : 0000 : 4 : 19 : 0 : 0]
단오 : [1 : 0 : 0000 : 5 : 5 : 0 : 0]
하지 : [1 : 1 : 0000 : 6 : 22 : 0 : 0]
개천절 : [1 : 1 : 0000 : 10 : 3 : 0 : 0]
근로자의 날 : [1 : 1 : 0000 : 5 : 0 : 1 : 1]
```

Fig. 3. 특정 어휘사전.

3. 정규화 규칙

시간 표현 어구는 정규화 규칙을 사용해서 절대시간(YYYYMMDD)으로 표현된다.

Table 5는 시간 정보를 정규화 하기 위한 방법으로 각 어휘별 의미에 따라 발행일 기준으로 생성한 정규화 규칙을 보여준다. '현재년도', '현재월', '현재일'은 기사가 발행된 발행일자를 기준으로 발행 년월일이 현재 년월일로 지정된다. 기간 정보는 사건이 언제 시작하여, 언제 종결되었는지를 추출하는 것으로 기간 표시에 사용된 어휘에 따라 정규화된 시간정보로 표기를 하게 된다. 기간의 시작임을 추측하게 하는 단어와 사용된 정규화된 시간은 시작 시간으로, 기간의 종결임을 추측하게 하는 단어와 사용된 정규화된 시간은 종결 시간으로 표기하게 된다. 범위정보는 해당 어휘가 가지는 의미에 따라 기간정보로 정규화된다.

4. 사건 동사 연결규칙

TDT 시스템에서 사건과 관련된 시간정보를 획득하기 위해서는 추출된 시간정보를 사건 동사와 연결하는 과정이 필요하다. 즉 문장 내의 동사는 사건에 대한 서술을 담당하기 때문에, 시간정보와 매핑된 동사는 사건을 탐지하고 추적하는 과정에서 사건의 선후관계를 규명하는데 중요한 자질로 사용된다. 문장 i에 대한 매핑 Mi=<T, S>이며, 여기서 T는 추출된 시간 표현이고 매핑 집합 S는 다음과 같다.

$$S=\{V_1, V_2, ..., V_n\}$$

Table 5. 시간 정규화 규칙

시 간	정규화 방법
과 거	작년, 지난해, 전년 : 현재년도 - 1
	지난달 : 현재월 - 1
	구랍 : 지난해 12월
	어제, 전날 : 현재일 - 1
현 재	올해, 금년, 올 : 현재년도
	이달 : 현재월
	이날, 오늘, 당일 : 현재일
미 래	내년 : 현재년도 + 1
	후년 : 현재년도 + 2
	내달, 다음달 : 현재월 + 1
	내일 : 현재일 + 1
기 간	월초 : 현재월 : 1일~현재월 : 10일
	상반기 : 현재년도 1월~현재년도 6월

여기서 V_i는 문장내 사용된 동사이다. 매핑 집합은 하나의 시간표현과 매핑될 수 있는 동사들의 집합으로 문장 내에 또 다른 시간표현이 나오기 전에 나타나는 동사들로 구성된다. 문장에서 하나의 시간표현만 나타난 경우 문장 내에서 사용된 모든 동사들이 하나의 매핑집합을 구성한다.

매핑 집합이 완성되면 다음과 같은 규칙에 의해 시간과 연결할 동사를 매핑 집합에서 선택하게 된다. 첫째, 시간과 동사의 위치정보를 고려하여 가장 인접한 시간과 동사를 매핑한다. 둘째, 관형격으로 사용된 동사와 종결형 동사가 연속되어 나타나는 경우 관형격 동사는 매핑 대상에서 제외된다. 예를 들면 "지난 3일 … 실시해온 조사에서 밝혔다"라는 문장의 경우 '지난 3일'이라는 시간과 '밝혔다'라는 동사를 연결하기 위해 '실시해온'이라는 관형격 동사를 매핑 대상에서 제외하였다. 셋째, 시간표현이 동사보다 많은 경우 마지막에 선택된 동사에 모든 시간정보를 매핑한다.

5. 추출과정

본 논문에서 제안하는 시간정보 추출 시스템은 문서 내에서 의미있는 시간정보를 포함한 문장을 대상으로 시간정보를 추출하고 이를 정규화한 후 연관성 있는 동사와 연결하여 사건의 탐지 및 추적에 사용할 수 있는 자질을 제공하는 것을 목표로 한다.

Fig. 4는 시간정보 추출 시스템의 진행 과정을 보여주고 있다. 첫째. 문서 집합을 대상으로 품사태깅을 수행하고, 둘째, 구축된 FSA 패턴과 품사별 어휘정보를 이용하여 시간을 표현하고 있는 문장을 추출한다. 셋째, 추출된 문장을 대상으로 시간 표현에 사용된 어휘 별 정규화 규칙을 적용하여 정규화를 수행하고, 넷째, 사건동사 연결규칙을 적용하여 문장에 나타나는 시간과 동사를 연결한다. 각 단계별 처리내용은 다음과 같다.

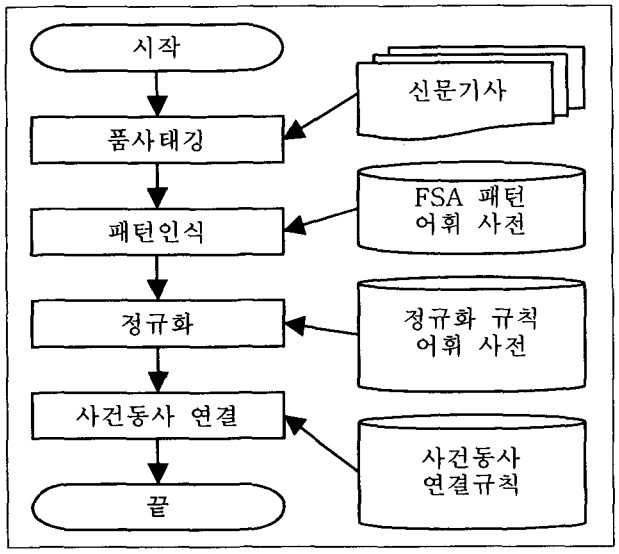

Fig. 4. 시간정보 추출과정.

1) 품사태깅

대상문서의 구조정보를 분석하기 위해 제일 먼저 품사태깅 과정을 수행한다. 품사태깅 과정은 시간 표현 어구를 추출하기 위한 첫 과정으로 태깅된 품사의 패턴과 품사별 어휘목록을 사용하게 되므로 매우 중요한 과정이다. 품사 태깅을 하기 위해서 고려대에서 개발한 품사 태거를 사용하였다.[15] 시간표현을 위한 어구는 '보통일반명사'(NNCG)로 태깅되어 그 자체로 시간을 표현하는 경우도 있지만, 보조사(PX)나 수관형사(DU)는 다른 시간 관련 태그와 결합해서 시간 표현이 가능한 경우도 있다. 이 밖에도 "지난"(동사+관형어말어미)과 같이 동사와 관형격 조사가 결합하여 시간 명사를 수식하는 경우와 "1.4분기"(숫자+기호(.)+숫자)와 같이 특수 기호를 포함한 시간 표현도 추출처리 대상에 포함시켰다.

2) 패턴인식

시각과 기간별 FSA와 어휘 사전을 사용하여 시간 표현 어구를 추출한다. 일차적으로 FSA를 사용하여 시간 표현의 가능성이 있는 어구를 선택하고 어휘 사전을 사용하여 최종적으로 시간 표현 어구인지를 결정한다.

3) 정규화

추출된 시간정보에서 시점과 기간을 'YYMMDD' 형식으로 표현하는 과정이다. 문장을 탐색하면서 시간정보가 나오면 이 형식으로 변환하고 이때 표시되지 않는 단위는 'X'로 표기한 후 기록한다. 연속되어 나타나는 시간정보에서 보다 작은 시간정보가 나타나면 기록된 'YYYYMMDD'에서 수정하여 기록한 후 관련 어구가 끝나면 이를 [시작시점 : 종료시점]으로 정규화하여 기록한다. 특정일과 관련된 어휘의 경우는 어휘 사전을 참조하여 절대시간으로 직접 연결하며, 이외의 시간은 정규화 규칙에 의하여 변환하게 된다.

4) 사건 관련 동사 연결

문장 내에 발생한 시간정보와 사건을 서술하기 위해 사용된 동사를 연결하는 작업을 수행한다. 본 연구에서는 문장의 구문 분석을 하지 않고, 단순한 규칙에 의하여 시간과 문장 내 동사를 연결하였다. 이 과정을 통해 문장 내에서 표현된 시간정보와 사건과의 연관성이 규명된다. 또한

다수의 시간정보가 하나의 기사 내에서 나타난 경우 해당 사건과 관련된 시간을 규명하기 위해서도 사용된다.

평 가(4장)

본 논문에서는 제안된 시간 정보추출 시스템을 평가하기 위해 다음 세가지 실험을 실시하였다.

[실험 1] 전체 시스템 성능 평가

[실험 2] 단계별 성능 평가

[실험 3] 다른 도메인 데이터 적용 실험

실험에 사용된 데이터는 한국경제신문과 한국일보 데이터를 내용별로 분류한 후, 데이터 구축에 사용된 범주와 그렇지 않은 범주를 대상으로 실험을 하였다. 시스템의 성능을 평가하기 위해 사람이 직접 각각의 테스트 데이터에 시점 및 기간, 정규화, 동사와의 매핑 결과를 추출한 결과와 시스템이 자동으로 추출한 결과를 비교 실험하였다. 다른 도메인에 대한 유용성 테스트를 위해서 사건/사고 도메인 중 교통사고와 관련된 데이터를 대상으로 실험을 하였다.

1. 전체 시스템 성능 평가

실험에 사용된 도메인에 해당하는 기사들은 10개의 범주로 구분할 수 있으며 시간 표현 패턴, 어휘 사전, 정규화 규칙, 사건 관련 동사 연결 규칙은 5개의 범주에 해당되는 기사를 대상으로 생성되었다.

Table 6은 실험에 사용된 문서의 분석 결과로 뉴스기사에서 시간정보를 포함하고 있는 문장의 수를 보여준다. 실험 데이터 분석은 패턴과 어휘 사전, 규칙들이 생성된 범주에 포함된 문서와 다른 범주에 포함된 문서로 구분하여 분석하였다. 대부분의 기사는 기사당 0~2문장이 시간정보를 포함한 문장으로 분석되었다.

Table 7는 시스템의 모든 단계를 거쳐 최종적으로 정규화된 시간정보와 동사의 매핑의 재현율과 정확도를 보여준다. 위 실험을 살펴보면 동일 범주 내 문서의 실험 결과가 다른 범주 문서의 실험 결과보다 좋게 나왔다. 이것은 다른 범주에 해당하는 문서의 경우 시간정보를 표현하기 위해 사용된 어휘나 패턴정보가 부족하기 때문이다. 각 단계별 실험을 통해 어휘사전이나 패턴이 전체시스템의 성능에 미치는 영향을 측정하였다.

Table 6. 동일 도메인 내 실험 데이터 분석

분류	동일 범주		다른 범주	
	범 주	문 장	범 주	문 장
범주 별 문장 수	도 산	27	가 계	53
	부동산	25	보 험	27
	수 입	42	수 출	55
	재 정	29	실적활동	41
	증 권	53	은 행	22
	계	176	계	198
계	374			

Table 7. 시간 정보 추출 실험 결과

항 목	동일 범주	다른 범주
총 문장 수	176	198
총 시간정보 수	270	347
추출된 시간정보 수	265	361
일치된 개수	174	179
재현율(recall)	0.644	0.516
정확도(precision)	0.657	0.496

2. 단계별 성능 평가

시스템의 단계를 다음과 같이 3단계로 구분하여 실험하였다. 첫째, 패턴 및 어휘 사전을 이용하여 문장에서 시간정보를 추출하는 단계, 둘째, 추출된 시간정보를 정규화 하는 단계, 마지막으로 정규화된 시간 정보와 관련 동사를 매핑하는 단계이다.

Table 8은 시간정보를 포함하고 있다고 추정되는 문장을 사람과 시스템이 추출한 후 그 재현율과 정확도를 실험하였다. 시간정보 추출 오류의 유형은 다음과 같이 5가지로 구분할 수 있다. 첫째, 띄어쓰기 오류로써 문장의 띄어쓰기가 잘못되어 품사태깅에 잘못된 결과가 나타나는 경우, 둘째, 품사 태깅의 오류로 인해 문장 내 단어의 구분이나 형태소 분석이 잘못된 경우, 셋째, 어휘 사전이 미등록된 단어가 시간정보를 표현하는데 사용된 경우, 넷째, 시간정보 패턴에 포함되지 않은 새로운 패턴이 발생하는 경우, 마지막으로 기간을 의미하는 단어가 단독으로 사용되는 경우 기간의 모호함을 발생시켜 제외하도록 설계되었기 때문

Table 8. 시간정보 추출실험 결과

항 목	동일 범주	다른 범주
총 문장 수	176	198
총시간정보 수	270	347
추출된 시간정보 수	265	361
일치된 개수	237	305
재현율(recall)	0.878	0.879
정확도(precision)	0.894	0.845

Table 9. 시간정보 정규화 실험 결과

항 목	동일 범주	다른 범주
총 정규화 수	237	305
일치된 개수	218	270
정확도(precision)	0.920	0.885

에 추출에서 제외된 경우이다.

Table 9는 시간 정보의 정규화 단계의 정확도를 보여준다. 정규화의 경우 문장 내 절대시간이 표현된 경우는 비교적 정확하고 정규화 된 반면, 상대적 시간표현 어구가 사용된 경우는 정규화에 오류가 있었다. 오류의 유형을 살펴보면 다음과 같다. 정규화 규칙이 없어서 잘못된 정규화를 수행한 경우, 문서의 의미를 파악하지 않기 때문에 과거 시점을 다룬 내용임에도 불구하고 문서의 생성 날짜만을 참조하여 정규화한 경우가 있으며, 내용 자체에서 정규화하기 모호한 문장도 이에 포함된다.

Table 10은 동사와 시간정보의 매핑에 대한 실험 결과를 보여준다. 본 실험에서는 동사의 시제와 태에 대한 고려를 하지 않았기 때문에 동사의 선후 관계를 고려하지 못하였고, 따라서 같은 문장에서 복수 개의 동사와 시간정보가 발생하는 경우 이에 대한 연결이 쉽지 않았다.

3. 다른 도메인 대상 실험

제안된 방법들의 확장성을 판단하기 위해서 다른 도메인의 문서를 대상으로 실험을 수행하였다. 패턴과 어휘 사전, 규칙들이 생성된 범주를 포함하는 경제 도메인과 교통사고 관련 기사들을 포함한 다른 도메인을 대상으로 비교 실험을 하였다. 즉 경제도메인에서 구축된 패턴과 어휘 목록을 사용하여 교통사고 도메인에 적용함으로써 구축된 패턴과 어휘 사전의 일반성을 평가하였다.

Table 11에 의하면 다른 도메인, 즉 교통사고 관련 기사들을 대상으로 시험한 결과 경제 도메인보다 좋음을 알 수 있다. 각 단계별 정확도를 분석한 결과 패턴 정보를 이용해서 시간정보를 포함한 문장의 추출 정확성과 동사의 매핑 정확성이 경제 도메인에 비해 높게 나타났다. 이와 같은 결과가 나온 이유는 경제 도메인의 경우 하나의 문장에 복수개의 시간정보가 나온 문장이 많은 반면, 교통사고 도메인의 경우 비교적 일관적이고 단순한 문장 패턴으로 기사가 작성되었기 때문이다. 또한 한 문장에서 표현하고자 하는 사건과 관련된 시간정보가 단일하게 나타나서 동사와 시간정보 매핑의 오류가 적었다.

Table 10. 동사 매핑 실험 결과

항 목	동일 범주	다른 범주
총 문장 수	218	270
일치된 개수	174	179
정확율(precision)	0.798	0.663

Table 11. 다른 도메인 대상 실험

항 목	동일 도메인	다른 도메인
총 문장 수	176	77
총 시간정보 수	270	81
추출된 시간정보 수	265	86
일치된 개수	174	61
재현율(recall)	0.644	0.753
정확도(precision)	0.657	0.709

결 론(5장)

본 논문에서는 뉴스기사나 방송기사를 대상으로 연구되는 사건 탐지/추적에 사용되는 중요한 자질 중 하나인 사건 관련 시간정보를 추출하는데 있어서 기존의 기사생성일과 사건발생일등에 차이가 있는 것을 고려하여 기사에 표기된 실제 시간정보를 자동 추출하는 시스템을 개발하는데 목표를 두었다. 제안된 시스템은 문서내의 시간 정보를 추출하기 위해 시간 정보 패턴과 어휘 사전, 시간 정규화 규칙, 관련 동사 매핑 규칙을 구축하고 이를 사용하여 시간정보를 추출하고 정규화하였다. 이렇게 추출된 시간 정보는 사건의 탐지/추적 분야 외에도 질의 응답, 정보 추출, 자동 요약, 기계 번역 분야에도 사용될 수 있다.

본 연구에서는 문서 내에서 절대 시간으로 표현된 시간 정보뿐만 아니라, 상대적으로 표현된 시간 정보도 같이 추출할 수 있도록 어휘 사전을 구축한 후, 패턴정보와 같이 사용하여 정확도를 높일 수 있었다. 시스템의 결과를 평가하기 위해서 데이터 구축에 사용된 문서와 같은 범주에 포함된 문서와 다른 범주에 포함된 문서를 대상으로 비교 실험하였고, 또한 다른 도메인에 해당되는 문서를 대상으로 비교적 좋은 결과를 얻음으로써 다른 도메인에도 사용할 수 있음을 알 수 있었다.

제안된 시간정보 추출 시스템은 품사태깅을 기반으로 구축되었고, 문서의 의미를 추론하는 과정이 없기 때문에 품사 태깅의 오류나 기사의 발생일과 기사 내 언급된 시간정보의 차이가 발생할 경우, 상대정보로 시간이 표현된 경우, 이를 정규화하는 과정에서 오류가 발생할 수 있다. 또한 동사의 시제나 태에 대한 고려가 없어 동사와 시간정보의 매핑 오류도 발생할 수 있다.

기사의 발생일 정보와 사건별 추출된 시간정보를 어떻게 활용해야 사건 탐지/추적의 정확성을 높일 수 있는지에 대한 방안은 지속적으로 연구되어야 한다. 또한 문서의 발표 시점 및 지역 문맥만을 가지고 시간 정보를 정규화 했던 것에서 벗어나, 문서 내의 시간의 흐름과 문장에 내재된 의미를 추론하는 과정을 도입하여 상대 시간을 정규화 하는

규칙의 개선이 필요하다. 또한 품사 태깅 정보 이외에도 동사의 시제나 태를 고려하여 시간정보 추출 시스템의 정확성을 높이는 것이 필요하다.

REFERENCES

1) Allan J, Papka R, Lavrenko V(1998) : *"On-line new event detection and tracking," Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp37-45*
2) Allan J, Carbonell JG, Doddington G, Yamron J, Yang Y(1998) : *"Topic Detection and Tracking Pilot Study Final Reprot," In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Feb*
3) Yang Y, Pierce T, Carbonell J(1998) : *"A Study on Retrospective and On-Line Event Detection," Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp28-36*
4) van Mulbregt P, Carp I, Gillick L, Lowe S, Yamron J(1998) : *"Text Segmentation and Topic Tracking on Broadcast News Via A Hidden Markov Model Approach," ICSLP98, Volume VI, pp2519-2522*
5) Yang Y, Carbonell J, Brown RD, Pierce T, Archibald BT, Liu X (1999) : *"Learning Approaches for Detection and Tracking News Events," IEEE Intelligent Systems, 14 (4) : 32-43, July/August Special Issue on Applications of Intelligent Information Retrieval*
6) Papka R, Allan J, Lavrenko V(1999) : *"UMASS approaches to detection and tracking at TDT2," In Proceedings of the TDT-99 workshop. NIST*
7) Sundheim B, Chinchor N(1995) : *"Named Entity Task Definition, Version 2.0, 31 May 95," Proc. of the 6th Message Understanding Conference (MUC-6), pp319-332, Morgan Kaufman Publishers, Inc*
8) Chinchor N(1998) : *"MUC-7 Information Extraction Task Definition, Version 5.1, 23 July 1998," Proc. of the 7th Message Understanding Conference (MUC-7)*
9) Ferro L, Mani I, Sundheim B, Wilson G(2001) : *"TIDES Temporal Annotation Guidelines," MITRE Technical Report Version 1.0.2, June*
10) Alice GB, Meulen T(1995) : *"Representing Time in Natural Language," MIT Press, Cambridge, Massachusetts*
11) Sowa JF(1984) : *"Conceptual Structures," Addison Wesley, Reading, Massachusetts*
12) Maiocchi R, Pernici B, Barbic F(1992) : *"Automatic Deduction of Temporal Information," ACM Transactions on Database Systems, Vol. 17, No. 4, pp647-688, December*
13) Moulin B(1997) : *"Temporal Contexts for Discourse Representation : An Extension of the Conceptual Graph Approach," Artificial Intelligence, 7 : pp227-255*
14) Juntae Yon, Yoonkwan Kim, Mansuk Song(2000) : *"Identifying Temporal Expression and its Syntactic Role Using FST and Lexical Data from Corpus," Colling*
15) Jin-Dong Kim, Heui-Seok Lim, Sang-Zoo Lee, Hae-Chang Rim (1998) : *"Twoply Hidden Markov Model : A Korean POS tagging Model Based on Morpheme-unit with Word-unit Context," Computer Processing of Oriental Languages, Vol. 11, No. 3, pp277-290*