JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 12, pp. 41-49, December 2022
https://doi.org/10.9708/jksci.2022.27.12.041

Rule-based Normalization of Relative Temporal Information

Young-Seob Jeong*, Chaegyun Lim**, SeungDong Lee*, Medard Edmund Mswahili*,
Goodwill Erasmo Ndomba*, Ho-Jin Choi**

*Professor, Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
**Student, School of Computing, KAIST, Daejeon, Korea
*Student, Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
*Student, Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
*Student, Department of Computer Engineering, Chungbuk National University, Cheongju, Korea
**Professor, School of Computing, KAIST, Daejeon, Korea

[Abstract]

Documents often contain relative time expressions, and it is important to define a schema of the relative time information and develop a system that extracts such information from corpus. In this study, to deal with the relative time expressions, we propose seven additional attributes of timex3: year, month, day, week, hour, minute, and second. We propose a way to represent normalized values of the relative time expressions such as before, after, and count, and also design a set of rules to extract the relative time information from texts. With a new corpus constructed using the new attributes that consists of dialog, news, and history documents, we observed that our rule-set generally achieved 70% accuracy on the 1,041 documents. Especially, with the most frequently appeared attributes such as year, day, and week, we got higher accuracies compared to other attributes. The results of this study, our proposed timex3 attributes and the rule-set, will be useful in the development of services such as question-answer systems and chatbots.

▸ **Key words**: Time information, Timex3, Relative temporal information, Normalization, Information extraction, Rule-based

• First Author: Young-Seob Jeong, Corresponding Author: Ho-Jin Choi
*Young-Seob Jeong (ysjay@chungbuk.ac.kr), Department of Computer Engineering, Chungbuk National University
**Chaegyun Lim (rayote@kaist.ac.kr), School of Computing, KAIST
*SeungDong Lee (sdlee130@naver.com), Department of Computer Engineering, Chungbuk National University
*Medard Edmund Mswahili (medardedmund25@chungbuk.ac.kr), Department of Computer Engineering, Chungbuk National University
*Goodwill Erasmo Ndomba (goodwillndomba23@gmail.com), Department of Computer Engineering, Chungbuk National University
**Ho-Jin Choi (hojinc@kaist.ac.kr), School of Computing, KAIST
• Received: 2022. 11. 14, Revised: 2022. 11. 29, Accepted: 2022. 12. 05.
• Young-Seob Jeong, Chaegyun Lim equally contributed to this paper.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

문서에는 상대적인 시간표현이 포함되어있으며, 이를 다루기 위한 시간표현 체계를 구축하고 상대시간정보를 추출하는 모델을 개발하는 것은 중요하다. 본 연구에서는 timex3 태그에서 상대적인 시간표현의 정규값을 담을 수 있도록 하기 위해 연, 월, 일, 주, 시, 분, 초 단위에 대하여 총 7가지의 새로운 속성을 새롭게 제시하였으며, 이전, 이후, 번째 등에 대한 정규값의 기술방법도 함께 제시하였다. 또한, 새롭게 추가된 속성들의 정규값을 추출하는 규칙 모음을 설계하였다. 추가된 속성들을 바탕으로 구축한 데이터셋은 일상대화, 뉴스, 역사와 관련된 총 1,041개의 문서를 포함하고 있으며, 본 연구에서 설계한 규칙 모음을 전체 데이터셋에 대하여 적용하여 전반적으로 70% 이상의 정확도를 보이는 것을 확인하였다. 특히, 데이터셋에 자주 등장한 상대시간표현인 year, day, week 속성에 대한 성능이 비교적 높은 것을 확인할 수 있었다. 본 연구의 결과물인 추가적인 timex3 속성과 규칙기반 모델은 질의응답시스템, 챗봇 등의 서비스 개발에 유용하게 활용될 수 있을 것이다.

▸ **주제어**: 시간정보, Timex3, 상대시간정보, 정규화, 정보추출, 규칙기반

I. Introduction

일상생활에서 접하는 자연언어 텍스트 데이터를 분석하여 발화자의 의도 또는 문장의미를 이해하는 모델을 개발하는 것은 여러 가지 서비스(예: 인공지능형 비서[1,2], 질의응답 시스템[3] 등)의 실현을 가능하게 하였다. 자연어 텍스트에는 사건, 시간 표현, 공간 표현, 의미역 등과 같은 다양한 정보가 포함되어 있으며, 이러한 정보를 정확하게 추출하는 것은 모델의 문장 이해력 향상에 긍정적인 영향을 주게 된다. 가령, '내일 비 와?'라는 문장에서 '내일'이라는 시간 표현에 대한 정보를 정확히 추출하지 못한다면 모델은 '오늘은 비가 옵니다'와 같은 엉뚱한 답변을 제공하게 될 것이다. '내일' 시간 표현에 대하여 정보를 추출하는 것은 2가지 과정을 내포한 것으로 볼 수 있는데, 첫 번째는 '내일'이 시간 표현인지 여부를 예측하는 것이다. 이는 개체명 인식 (Named entity recognition) task 또는 슬롯 (Slot) 추출 task와 밀접하며, 추출된 '내일'은 개체명 (Named entity) 또는 슬롯 (Slot)이라고 부를 수 있다. 두 번째는 '내일'이 하루 뒤라는 의미까지 파악하는 작업이며, 이를 시간 정규화 작업이라고 한다. '내일'이 날짜에 대한 시간 표현이라는 것을 찾아내는 것도 중요하지만, 하루 뒤에 해당되는 의미인 것을 알아내는 것이 서비스 관점에서 더욱 중요하다는 것은 의심의 여지가 없다. 하지만, 인공지능 기술이 나날이 발전하고 있는 작금의 상황에도 불구하고 시간 정보 정규화 작업에는 최신 딥러닝 기술들이 성공적으로 적용되지 못하고 있다.

시간 정보를 추출하는 모델을 개발하기에 앞서 선행되어야 할 것은 시간 정보를 정의하고 annotation scheme을 설계하는 것이다. ISO-TimeML [4]을 비롯하여 한국어를 위해 설계된 annotation scheme [5,6] 들이 제시되었으며, 이 scheme 들을 사용하여 시간 정보 데이터셋이 구축되기도 하였다. 하지만, 지금까지 제시되었던 annotation scheme 에서는 '상대적인 시간 정보'를 담은 표현을 위한 정규화 방법이 포함되지 않았다는 한계점이 있다. 예를 들어, 오늘이 2022년 12월 24일이라는 가정하에 '내일 비 와?' 라는 발화문장에 대하여 시간 정보를 태깅한다면 아래와 같이 timex3 태그를 만들 수 있을 것이다. 하지만, timex3 태그의 정규화된 값을 가지는 'value' 속성에는 2022-12-25 라는 값이 들어가게 되는데, 이는 Document Creation Time (DCT)이 2022년 12월 24일이라는 것을 전제로 내일 날짜를 계산한 결과이다. 이처럼 기존의 annotation scheme들은 timex3 태그의 value 속성에 '절대적인 시간 값'을 정규화하도록 하고 있는데, 만약 오늘 날짜를 모르는 상황일 경우에는 절대적인 시간 값을 얻는 것이 불가능하다는 한계점이 존재한다.

[DCT] 2022-12-24
[문장] 내일 비 와?
[TIMEX3 태그] text='내일', type='DATE', value='2022-12-25'

본 연구에서는 기존의 annotation scheme 에서는 다루지 않은 '상대적인 시간 값'의 정규값을 담을 수 있도록 하기 위해 timex3 태그에 대한 새로운 속성들을 제시한다.

또한, 새로운 속성을 사용하여 구축한 한국어 말뭉치를 대상으로 상대적인 시간 값을 추출하는 모델을 개발하였다.

II. Preliminaries

1. Korean TimeML

ISO-TimeML은 시간정보 annotation 스키마의 국제 표준이며, 이것이 포함하고 있는 주된 태그는 3가지(timex3, event, tlink)이다. Timex3 태그는 '12월 24일', '다음 주', '작년', '2달 전' 등과 같은 시간 표현에 대한 태그이며, ISO-TimeML에서 가장 중요한 태그라고 볼 수 있다. ISO-TimeML은 영어에 집중하여 설계된 측면이 있어서, 이를 언어에 특화시킨 스키마를 제시했던 과거 연구들이 있어왔다 [5-7]. 한국어를 위한 TimeML인 Korean TimeML [5] 이 등장한 이후 2016년에 이를 개선한 새로운 스키마가 등장하였는데 [6], 이는 timex3, event, tlink 태그의 일부 속성(attribute)들을 수정하거나, 한국어에 종종 등장하는 시간 표현들(예: 음력 시간표현, '초중반', '중후반' 등)을 다루기 위해 새로운 속성을 제시하기도 하였다.

기존 Korean TimeML을 바탕으로 데이터셋들이 구축되기도 하였으나, 해당 Korean TimeML은 '상대적인 시간표현'을 다루기 위한 방법을 포함하고 있지 않다는 점에서 여전히 한계가 있다. 예를 들어, '내일'이라는 시간 표현은 timex3 태그에 해당되며, 'value' 속성에는 절대적인 시간 정규값이 들어가게 된다. 예를 들어, 오늘이 2022년 12월 24일일 경우에는 '내일'의 절대적인 시간 정규값인 '2022-12-25'라는 값이 'value' 속성에 들어간다. 하지만, 만약 오늘 날짜가 주어져있지 않은 경우에는 '내일' 시간 표현에 대한 절대적인 정규값을 알 수가 없기 때문에 timex3 태그의 'value' 속성에는 'XXXX-XX-XX'와 같은 값이 들어가게 되며, '내일'이 '하루 뒤'에 해당된다는 상대적인 시간 정보를 표현할 방법이 없다. 이러한 한계점을 해결하기 위해 'day' 속성을 추가하고 '+1'이라는 상대적인 정규값(하루 뒤를 의미)을 담도록 할 수 있다. 본 연구에서는 'day' 속성 외에도 다양한 시간 단위(예: 연, 월 등)에 대한 상대적인 시간표현을 위한 속성들을 제시하며, 일상 대화에서 '다음달에 봐!', '내일 만나자' 등과 같은 상대적인 시간표현이 자주 등장하는 점을 미루어볼 때 본 연구의 결과물이 다양한 서비스 개발에 유용하게 사용될 것으로 기대된다.

2. TIMEX3 Normalization

2.1 Task description

Timex3 태그 추출 task는 주어진 텍스트에서 시간 표현에 해당되는 텍스트 일부를 추출하는 것이며, 이는 기본적으로 2단계를 거친다. 첫 번째 단계는 주어진 텍스트에서 timex3 태그에 해당되는 범위(extent)를 찾는 것이며, 두 번째 단계는 timex3 태그의 속성 값들을 찾는 것이다. Timex3 태그의 범위(extent)를 찾을 때는 주로 sequence labelling 문제를 푸는 방식을 사용할 수 있는데, 그림 1 은 글자 단위의 BIO notation을 사용하여 timex3 태그의 범위(extent)를 찾는 예시를 보여주고 있다. 시간 표현에 해당되는 'yesterday'에는 DATE 라는 값으로 태깅되어 있는데, 이는 timex3 태그의 'type' 속성이 가질 수 있는 4가지 값(DATE, TIME, DURATION, SET) 중의 한 가지에 해당된다. 즉, BIO notation을 사용할 때, 단순히 timex3 태그의 범위(extent)만 찾는 것이 아니라 'type' 속성의 값도 함께 추출하게 되는 것이며, 이는 개체명 인식 (named entity recognition) task에서 취하는 sequence labelling 방식과 동일하다. 범위(extent)를 찾아낸 timex3 태그는 'yesterday', 즉 '하루 전'을 의미하므로, 오늘 날짜가 2022년 12월 24일일 경우에는 'value' 속성에 '2022-12-24'라는 절대시간 정규값을 추출할 수 있게 된다.

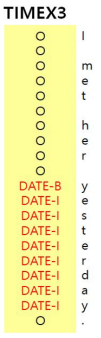


Fig. 1. Character-level sequence labelling of TIMEX3 tag

2.2 Timex3 normalization method

인터넷 상의 텍스트 데이터 양이 폭발적으로 증가함에 따라, 자연언어처리 (natural language processing) 분야에서의 인공지능 기술의 중요성이 커지고 있으며, 특히 대량의 미태깅 데이터(unlabelled data)를 학습한 거대 언어모델을 개발하는 시도가 계속되고 있다. 이러한 거대 언어모델은 대량의 텍스트 데이터에 감춰진 언어적 패턴을 자동으로 학습하여, 여러 가지 자연언어처리 분야 task들(예: 문장 분류 [8,9], 개체명 인식 [10], 텍스트 생성 [11] 등)과 다른 종류의 데이터와 혼합된 멀티모달 task들에서 높은 성능향상을 보여주고 있다 [12,13].

거대 언어모델들은 주로 Transformer [14]의 self-attention 구조에 기반하고 있는데, 이는 입력으로 주어진 텍스트 구성요소인 토큰들 간의 모든 가능한 관계 패턴을 학습할 수 있으며, 층(layer)이 쌓여감에 따라 더욱 높은 수준의 패턴을 학습할 수 있게 된다. 그럼에도 불구하고 이러한 거대 언어모델들이 여전히 풀지 못하는 task들이 있다. 예를 들어, 수학 문제를 푸는 거대 언어모델 Minerva [15] 는 수학 문제의 수치값들이 바뀌면 성능의 편차가 커지는 것을 발견했으며, 이는 거대 언어모델들이 수치형 값의 정규화에 약하다는 것을 보여준다고 할 수 있고, 시간 표현에 대한 timex3 태그의 속성 정규값 예측에도 거대 언어모델들은 한계가 있다는 것을 짐작할 수 있다. 과거 연구들에서도 timex3 태그의 범위(extent) 또는 'type'을 예측할 때는 기계학습 또는 딥러닝 기반의 소위 data-driven approach 방식을 채택했으나, 'value' 속성 등과 같은 정규값 예측에는 규칙에 기반하여 예측하였다 [17].

본 연구에서는 상대적인 시간 표현을 다루기 위해 timex3 태그의 추가적인 속성들을 정의하고, 규칙 세트를 기반으로 상대적인 시간 정규값을 예측하였다.

III. The Proposed Scheme

1. TIMEX3 Attributes for Relative Time Expressions

본 연구에서는 상대적인 시간표현에 대한 정규화를 위해 timex3 태그에 대하여 아래 Table 1과 같이 7개의 추가적인 속성들을 제시한다. year, month, day는 각각 연, 월, 일에 대한 상대적인 시간표현의 정규값을 담는 용도로 사용되며, hour, minute, second는 각각 시, 분, 초에 대한 상대적인 시간표현의 정규값을 담는다. week은 '주'에 대한 상대적인 시간표현 정규값을 가진다. 이 속성들은 절대적인 시간값을 담기 위한 것이 아니다. 예를 들어, '2010년 3월'이라는 시간표현은 상대 시간표현이 아니며, 이 시간표현의 정규값은 value='2010-03'과 같이 value 속성에 기입하며, year와 month 속성의 값은 존재하지 않게 된다. 반면, '내년 3월'이라는 시간표현에 대해서는 year='+1', value='XXXX-03' 과 같이 (DCT가 주어지지 않았다고 가정할 경우에는 내년이 몇 년도인지 알 수 없으므로 XXXX가 됨) 정규값을 기입할 수 있으며, 이 때 year 속성의 '+1'은 1년 뒤를 의미하는 상대적인 시간 정규값이다. 본 연구에서 새롭게 제시한 속성들은 -N, 0, +N, 그리고 ~N 형태(N은 자연수 또는 실수)의 값을 가질 수 있는데, 0은 현재의 시간을 나타내며 -N은 과거, +N은 미래, 그리고 ~N은 차수 혹은 임의의 번째를 나타내게 된다. 즉, year='0'이 의미하는 것은 '올해'를 의미하고 month='0'이 의미하는 것은 '이번달'이 되며, year='+1'은 '내년', month='+1'은 '다음달', day='~3'은 '3일차'를 의미하는 정규값이다. 여기서 '~N' 형태의 정규값은 연차, 주차 등과 같이 year, month, day 등에서 임의의 차수를 나타내는 시간 표현에 대한 값인데, '오늘은 행사 3번째 날이잖아'와 같은 문장에서는 '3번째 날'에 대하여 day='~3' 이라고 정규값을 지정한다. 이 경우, '행사'의 정확한 일시를 알지 못하는 경우라면 value='XXXX-XX-XX' 등과 같은 형태로 절대 시간값을 태깅하게 되며, 이는 value 속성만으로는 행사의 3번째 날이라는 중요한 시간 정보를 태깅하지 못한다는 의미이다.

Korean TimeML 2016 에 따르면, timex3 태그에는 value, freq, mod 등의 여러 가지 속성들이 있으며, 그 중에서 'value' 속성은 아래 예시와 같이 timex3 태그의 시간 표현이 가진 절대적인 정규값을 담게 된다. 이 때, 해당 텍스트가 만들어진 해가 2022년이라는 가정 하에 ISO-8601 [16]에 의거하여 value 속성에 2022-05-06 이라는 값을 넣게 된다.

[DCT] 2022-05-01
[Text] 학교에서 5월6일에 보자.
[TIMEX3 태그] text='5월6일', type='DATE', value='2022-05-06'

Table 1. Proposed TIMEX3 attributes for relative time expressions.

Name	Meaning
year	Relative value of year (e.g., 'last year': year="-1")
month	Relative value of month (e.g., 'this month': month="0")
day	Relative value of day (e.g., '3rd day': day="~3")
hour	Relative value of hour (e.g., '3 hours ago': hour="-3")
minute	Relative value of minute (e.g., 'after 1 minute': minute="+1")
second	Relative value of second (e.g., 'after five seconds': second="+5")
week	Relative value of week (e.g., '3 weeks later': week="+3")

절대적인 정규값을 담는 'value'는 시간 정보에 있어서 매우 중요한 속성인 것은 분명하나, 상대적인 시간 표현에 대해서도 절대적인 정규값만을 담게 되어있다는 점에서는 한계점이 있다. 아래 예시와 같이, 문장 생성 날짜가 주어져있지 않은 상황에서는 '내일'에 해당되는 timex3 태그의 value 속성의 정규값을 특정할 수 없게 된다. 만약 문서

생성 날짜가 2022년 12월 31일이면 value='2023-01-01'이 될 것이고, 문서 생성 날짜가 2022년 1월 1일이면 value='2022-01-02'가 될 것이다. 이처럼, 문서 생성 날짜에 따라서 value 속성의 값은 완전히 바뀔 수 있기 때문에 절대적인 시간 정규값만을 담는 value 속성은 '내일'이라는 시간 표현이 가진 정보를 온전히 담지 못하게 된다. 하지만, 본 연구에서 제안한 속성들 중의 하나인 day 속성에는 '+1'이라는 값을 담도록 함으로써 1일 뒤라는 상대적인 시간 정규값을 가질 수 있다.

[DCT] N/A (없음)
[Text] 학교에서 내일 보자.
[TIMEX3 태그] text='내일', type='DATE', value='XXXX-XX-XX', day='+1'

본 연구에서 제안한 속성들은 'value' 속성과 독립적으로 사용되어야 한다. 만약 문서 생성 날짜가 2022년으로 주어졌을 경우, 아래 예시와 같이 '내년'이라는 시간 표현의 절대적인 정규값을 value 속성에 담을 수 있다. 또한, year 속성에는 '내년'의 상대적인 정규값인 '+1'을 담는다. 즉, value 속성과 year 속성은 독립적으로 사용되기 때문에, 아래 예시의 timex3 태그에서는 value 속성의 2023년이라는 절대 정규값과 year 속성의 '+1'이라는 상대 정규값을 별개로 다루어야 한다. 이렇게 하는 이유는, ISO-TimeML을 비롯하여 그동안 제시되었던 annotation scheme에서 사용한 value 속성의 정의를 침해하지 않으면서도 상대적인 정규값을 다룰 수 있도록 하기 위함이다. 만약 '2022년 5월 3일'과 같이 명확하게 절대적인 시간 정규값을 알 수 있는 시간표현의 경우에는 value 속성에만 정규값을 기록하고 year, month, day 등과 같은 상대적인 시간값을 위한 속성들의 값은 비워두어야 한다.

[DCT] 2022-12-01
[Text] 내년에 보자.
[TIMEX3 태그] text='내년', type='DATE', value='2023' year='+1'

2. Normalization of TIMEX3 Attributes

본 연구에서 제시한 timex3 태그의 상대적인 정규값을 위해 추가 정의된 속성들에 대한 예측 task를 수행하기 위해 규칙 기반의 모델을 구축하였으며 그 규칙은 아래 Table 2, Table 3과 같다. Table 2는 year, month, day에 대한 규칙이고, Table 3은 hour, minute, second, week 에 대한 규칙이다. 여기서 N은 숫자 또는 기수 등의 형태일 수 있으며, 규칙에 의한 정규화된 값을 취하게 된다. 예외적으로, N이 '몇'이라는 텍스트인 경우에는 정규값을 'X'로 취한다. 예를 들어, '몇년차'라는 시간표현의 '몇'은 정확한 수치를 나타내는 표현이 아니므로 이에 대한 timex3 태그의 year 속성의 값은 "~X"가 된다. 상대 정보의 정규값을 위해 만든 규칙임에도 불구하고, 이 규칙을 활용함으로써 timex3 태그의 extent (텍스트 상에서의 태그 범위)와 type 속성도 함께 예측하였다. 모델이 잘못된 extent와 type 값을 예측할 경우에는 상대적인 정규값은 항상 잘못된 값을 예측하게 되므로, extent와 type의 값을 정확하게 생성함과 동시에 상대적인 시간 속성의 올바른 값도 추출하도록 규칙을 설계해야 할 것이다.

Table 2. List of rules for normalization of year, month, day, and week attributes. 'N' in rules denotes a sequence of numbers or a cardinal number, whereas 'N' in values indicates a normalized number.

Rules		Value
year	N년(앞|전)	year="-N"
	N년(뒤|후)	year="+N"
	내년|다음(해|년)	year="+1"
	작년|(저번|지난|이전)해	year="-1"
	내후년	year="+2"
	재작년	year="-2"
	이번(해|년)|올해	year="0"
	N년(차|째)	year="~N"
month	N개월(앞|전)	month="-N"
	N개월(뒤|후)	month="+N"
	(내|다음)달	month="+1"
	(저번|지난|이전)달	month="-1"
	다다음달	month="+2"
	저저번달	month="-2"
	이번달	month="0"
	N개월(차|째)	month="~N"
day	N일(앞|전)	day="-N"
	N일(뒤|후)	day="+N"
	내일|다음날	day="+1"
	어제|이전날	day="-1"
	(하루|이틀|사흘|나흘|닷새|엿새|이레|여드레|아흐레|열흘|...)뒤	순서대로 각각, day="+2", "+3", ...
	(하루|이틀|사흘|나흘|닷새|엿새|이레|여드레|아흐레|열흘|...)전	순서대로 각각, day="-2", "-3", ...
	모레|(사흘|글피)|나흘	순서대로 각각, day="-2", "+3", "+4"
	그그저께|그저께	순서대로 각각, day="-3", "-2"
	N일(차|째)	day="~N"
week	N주(앞|전)	week="-N"
	N주(뒤|후)	week="+N"
	(저번|이번|다음)주	순서대로 각각, week="-1", "0", "+1"
	N주(차|째)	week="~N"

Table 3. List of rules for normalization of hour, minute, second, and week attributes. 'N' in rules denotes a sequence of numbers or a cardinal number or a , whereas 'N' in values indicates a normalized number.

Rules		Value
hour	N시간(앞|전)	hour="-N"
	N시간(뒤|후)	hour="+N"
	N시간째	hour="~N"
minute	N분(앞|전)	minute="-N"
	N분(뒤|후)	minute="+N"
	N분째	minute="~N"
second	N초(앞|전)	second="-N"
	N초(뒤|후)	second="+N"
	N초째	second="~N"

Table 2와 Table 3의 규칙은 연, 월, 일, 주, 시 분, 초에 대하여 독립적으로 기술된 규칙이며, 실제로 이를 적용할 때에는 이 규칙들을 큰 시간 단위부터 작은 시간단위로 연쇄 적용한다. 예를 들어, 아래와 같이 '2년 3개월'이라는 시간 표현을 '2년'과 '3개월'이라는 개별적인 시간 표현으로 보지 않고, year에 대한 규칙과 month에 대한 규칙을 연쇄 적용함으로써 1개의 timex3 태그로서 생성하게 된다.

[DCT] N/A (없음)
[Text] 2년 3개월 뒤에 보자.
[TIMEX3 태그] text='2년 3개월 뒤', type='DATE', value='XXXX-XX', month='+3', year='+2'

IV. Experiments

Korean TimeML 2016을 기초로 하되 본 연구에서 제시한 추가적인 속성들을 바탕으로 구축한 데이터셋을 사용하여 실험하였다. 규칙 기반의 모델을 구축하였으므로 전체 데이터를 테스트 용도로 사용하였다. 데이터는 대화, 뉴스, 역사 분야에 대한 문서들을 포함하며, 뉴스의 경우에는 사회 및 경제 분야 뉴스 문서들이다. 대화는 총 5가지 주제(연애, 진로, 여행, 취미, 기타)에 대하여 1턴~5턴 사이의 발화 텍스트를 담고 있다. 데이터에 대한 명세는 Table 4에 요약되어있다. '# of attributes'는 상대적인 시간표현에 대한 속성인 year, month, day, week, hour, minute, second 의 값이 존재(예: +1, -3, 0, ~3 등)하는 태그의 개수이다.

데이터는 JSON 포맷으로 뉴스 문서, 대화 세션, 역사 문서 당 1개의 파일로 구성되었으며, 문서 생성시간, 수집 소스, 작성자(예: 뉴스 기자 등) 등에 대한 메타 데이터를 포함하고 있다. 대화 데이터의 '세션'은 최초 발화부터 임의의 턴 개수만큼 발화가 오고감으로써 대화가 끝나기까지를 의미한다. Fig. 2는 대화 세션에서 1개의 발화에 해당되는 데이터 샘플을 보여주고 있다.


```
{
  "_id": {
    "$oid": "633f9c32c8f08c6d66bccb0d"
  },
  "utterances": [
    {
      "event": [
        {
          "class": "OCCURRENCE",
          "comment": "",
          "extent": [
            15,
            18
          ],
          "id": "EVENT_S0_0",
          "subobj": "{"S":[1], "O":[3]}",
          "temp_s_id": 0,
          "text": "운영하"
        }
      ],
      "text": "음악 분수는 오늘 몇시부터 운영하나요?",
      "timex3": [
        {
          "calendar": "GREGORIAN",
          "comment": "",
          "day": "0",
          "extent": [
            7,
            12
          ],
          "hour": "",
          "id": "U0_T0",
          "minute": "",
          "mod": [],
          "month": "",
          "second": "",
          "temp_s_id": 0,
          "text": "오늘 몇시",
```


Fig. 2. Sample of dialog data file.

Table 4. Data statistics.

Item		Value
# of documents	Dialog	624
	News	174
	History	243
Avg. # of sentences (utterances) of a document	Dialog	5.78
	News	17.43
	History	8.11
Avg. # of words of a sentence	Dialog	9.29
	News	14.84
	History	13.65
# of TIMEX3 tags	Dialog	3,765
	News	1,800
	History	1,270
# of attributes	year	726
	month	165
	day	582
	week	209
	hour	8
	minute	6
	second	1

데이터에 annotation되어있는 timex3 태그들의 모음을 D_{gold}라고 정의하고, 텍스트 데이터에 대하여 모델을 적용하여 추출된 timex3 태그들의 모음을 D_{pred}라고 정의하도록 하자. D_{gold}에 존재하는 각 timex3 태그와 D_{pred}의 timex3 태그들 간에 extent를 비교하여 겹치는 태그 쌍을 구할 수 있으며, 이렇게 구한 모든 태그 쌍의 모음을 D_{pair}

= {(g, p) | $p \in D_{pred}$, $g \in D_{gold}$} 라고 하자. 각 태그 쌍 (g, p)에 대하여, g 태그의 year, month, day, week, hour, minute, second 속성의 값이 존재할 경우(예: +3, -1, 0, ~X 등), p 태그의 해당 속성에 기록된 예측값이 정확히 일치하는 경우에만 맞춘 것으로 간주하여 정확도를 계산하였다. 즉, 정확도는 속성의 값을 정확히 맞춘 개수를 D_{gold}에 존재하는 태그 개수로 나눈 값이다. 만약, i번째 $g_i \in D_{gold}$와 extent가 겹치는 $p_i \in D_{pred}$가 없는 경우에는, g_i에 대한 예측이 틀린 것으로 간주하였다. 즉, 본 연구에서 만든 규칙은 상대 시간표현에 대한 정규값을 찾기 위한 규칙이나, timex3 태그의 extent를 제대로 찾지 못한 경우는 상대 시간표현에 대한 정규값 예측도 틀린 것으로 간주한 것이다.

Table 5는 실험결과를 요약하고 있다. minute 속성에 대하여 50%, month에 대하여 약 61%를 달성한 점만 제외하면 대체적으로 70% 이상의 정확도를 달성했다. 데이터 개수가 상대적으로 많은 year, day, week 속성에 대한 정확도가 높게 나왔는데, 이는 이 속성들에 대한 시간 표현들이 상대적으로 문서에 자주 등장한다는 것을 의미하는 한편, 이 속성들이 텍스트에 등장할 때 주로 등장하는 패턴이 존재한다는 것을 의미한다. 가령, year 속성에 대해서는 '내년', '작년' 등과 같은 표현들이 주로 등장하고, day 는 '어제', '내일' 등의 표현이 주류를 이룬다.

Table 5. Accuracy of each attribute.

Attribute	Accuracy(%)
year	73.14%
month	61.21%
day	74.57%
week	91.39%
hour	75.00%
minute	50.00%
second	100.00%

Timex3 태그의 extent를 제대로 찾지 못한 경우에는 맞추지 못한 것으로 간주했다는 점을 감안하면, 전반적으로 70% 이상의 정확도를 거둔 규칙 기반 모델은 준수한 결과를 보였다고 할 수 있다. 특히, 한국어 시간표현에 대하여 종합적인 정보추출로써 수행되었던 연구에서 절대적인 시간값을 담는 'value' 속성에 대한 정규값 예측에 대한 recall 이 63.44%, F1 score 가 67.08% 였다는 점을 감안할 때 [17], 본 연구의 규칙 기반 모델에 의해 추출되는 상대적인 시간표현의 정규값 정확도는 준수하다고 볼 수 있다.

V. Conclusions

문서에 포함된 시간 표현에 대한 timex3 태그의 정규화된 값을 예측하는 것은 매우 중요하다. Timex3 태그에 대한 기존의 annotation scheme 에서는 '상대적인 시간 표현'(예: 내일, 작년, 일주일 뒤 등)에 대한 정규값을 담는 속성이 체계화되어있지 않았는데, 본 연구에서는 상대적인 시간 표현을 다루기 위해 timex3 태그의 추가적인 속성들을 제안하였다. 추가 속성들을 기반으로 구축된 데이터셋을 사용하여 상대적인 시간표현에 대한 속성들의 정규값을 예측하기 위한 규칙 기반 모델의 성능을 테스트하였으며, 전반적으로 70% 이상의 정확도를 달성하였다. 본 연구의 결과물은 질의응답시스템, 챗봇 등에 도움이 될 것으로 생각된다. 가령, 오늘 날짜가 주어져있지 않을 경우, "내일 밥 먹으러 가자"라는 문장에 대하여 기존의 annotation scheme에 기반하여 구축된 예측 모델은 timex3 (text='내일', value='XXXX-XX-XX', type='DATE',) 와 같은 형태로 결과를 주게 되는데, 이는 챗봇 서비스 관점에서는 좋은 응답문을 생성하기 위한 정보를 담지 못한다고 볼 수 있다. 반면, 본 연구의 결과물은 timex3(text='내일', value='XXXX-XX-XX', day='+1', type='DATE')와 같이 day 속성에 '1일 뒤'라는 정보를 담게 되므로, 챗봇이 '내일'의 절대적인 시간값을 유추하기 위해 외부 리소스를 활용하거나 사용자에게 재질의를 하는 등의 행위를 하도록 개발하는 것이 가능해진다.

본 연구에서 구축한 상대시간 정규값 예측 모델은 규칙을 기반으로 동작하기 때문에, 심화된 규칙을 설계하기 위해서 전문가의 지속적인 관리 및 작업이 요구된다는 점에서 한계가 있다. 본 연구에서 제시한 새로운 속성들을 바탕으로 양질의 데이터를 계속 누적한다면, 추후 딥러닝을 비롯한 data-driven 기법들을 적용하여 정규값을 예측하는 기술을 적용함으로써 높은 성능을 기대할 수 있을 것이다.

ACKNOWLEDGEMENT

This research was supported by Chungbuk National University Korea National University Development Project (2022)

REFERENCES

[1] Naver Clova, https://clova.ai/ko (accessed Oct. 24th, 2022)

[2] KT Giga gini, https://gigagenie.kt.com/ltemain.do (accessed Oct. 24th, 2022)

[3] IBM Watson, https://www.ibm.com/watson (accessed Oct. 24th, 2022)

[4] James Pustejovsky, Kiyong Lee, Harry Bunt and Laurent Romary, "ISO-TimeML: An International Standard for Semantic Annotation," Proceedings of the Seventh International Conference on Language Resources and Evaluation, pp. 394-397, Valletta, Malta, 2010.

[5] Hyun-Jo Yu, Hayeun Jang, Yumi Jo, Seungho Nam, Hyopil Shin, and Yoon-shin Kim, "The Korean TimeML: A Study of Event and Temporal Information in Korean Text," Language and Information, Vol. 15, No. 1, pp. 31-62, June 2011. 10.29403/LI.15.1.3

[6] Young-Seob Jeong, Won-Tae Joo, Hyun-Woo Do, Chae-Gyun Lim, Key-Sun Choi, and Ho-Jin Choi, "Korean TimeML and Korean TimeBank," Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 356-359, Portorož, Slovenia, 2016.

[7] Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof, "Annotating Events, Temporal Expressions and Relations in Italian: the It-Timeml Experience for the Ita-TimeBank," Proceedings of the 5th Linguistic Annotation Workshop, pp. 143-151, Portland, Oregon, USA, 2011.

[8] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld, "Pretrained Language Models for Sequential Sentence Classification," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 3693-3699, Hong Kong, China, 2019. 10.18653/v1/D19-1383

[9] Gati L. Martin, Medard E. Mswahili, Young-Seob Jeong, and Jiyoung Woo, "SwahBERT: Language Model of Swahili," Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.314-324, Seattle, Washington, 2022. 10.18653/v1/2022.naacl-main.23

[10] Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P. Xing, "Effective Use of Bidirectional Language Modeling for Transfer Learning in Biomedical Named Entity Recognition," Proceedings of the Machine Learning for Healthcare Conference, pp. 383-402, Palo Alto, California, 2018.

[11] Touseef Iqbal and Shaima Qureshi, "The survey: Text generation models in deep learning," Journal of King Saud University - Computer and Information Sciences, Vol. 34, No. 6, pp. 2515-2528, June 2022. 10.1016/j.jksuci.2020.04.001

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," arXiv:2205.11487, 2022.

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

[15] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra, "Solving Quantitative Reasoning Problems with Language Models," arXiv:2206.14858, 2022.

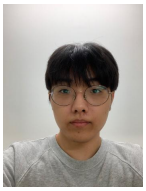
[16] ISO 8601, https://www.iso.org/iso-8601-date-and-time-format.html (accessed Oct. 24th, 2022)

[17] Young-Seob Jeong, Zae Myung Kim, Hyun-Woo Do, Chae-Gyun Lim, and Ho-Jin Choi, "Temporal Information Extraction from Korean Texts," Proceedings of the 19th SIGNLL Conference on Computational Natural Language Learning, Beijing, China, pp. 279-288, 2015. 10.18653/v1/K15-1028

Authors

Young-Seob Jeong received the M.S., Ph.D degree in computer science from KAIST, Daejeon, Korea, in 2012 and 2016, respectively. He is a faculty member of the department of computer engineering, Chungbuk National university, Cheong-ju city, Korea. His current research topics include malware detection using deep learning techniques, language models, healthcare system for patients, and pharmaceutics.

Chaegyun Lim received a B.S. in Medical Computer Science from Eulji University, Korea in 2011. He worked as a research assistant in the Department of Computer Science at KAIST, Korea, between 2011 and 2013. He obtained a M.Sc. in the Department of Computer Engineering at Kyung Hee University, Korea, in 2015. He is currently a PhD candidate in the School of Computing at Korea Advanced Institute of Science and Technology (KAIST), Korea. His research interests include temporal information extraction, multi-task learning, big data analysis and bioinformatics.



SeungDong Lee is an undergraduate student in the department of computer engineering, Chungbuk National university, Cheong-ju city, Korea. He will begin master course in the Computer Engineering department at Chungbuk National University, Cheong-ju city, South Korea. His research topics include machine learning, deep learning, and natural language processing.

Medard Edmund Mswahili received his M. Sc. Eng in Big Data Engineering, ICT convergence department from Soonchunhyang University, Asan, South Korea in 2022. He's currently pursuing his Ph.D. as a research assistant at Data Analysis and Artificial Intelligence Lab in the Computer Engineering department at Chungbuk National University, Cheong-ju city, South Korea. His research interest mainly focuses in Machine & Deep Learning in Pharmaceutical data analysis (drug discovery & development) and natural language processing.

Goodwill Erasmo Ndomba is a computer engineering Master's Degree student at Chungbuk National University, Cheong-ju city, Korea. He is a member of Data Analysis and Artificial Intelligent reasearch Labaratory, Computer Engineering department, Chungbuk National University, Cheong-ju city, South Korea. His current research area includes language models of natural language processing.

Ho-Jin Choi received a B.S. in Computer Engineering from Seoul National University, Korea, in 1982. He obtained an M.Sc. in Computing Software and Systems Design from Newcastle University, UK, in 1985 and a Ph.D. in Artificial Intelligence from Imperial College, London, UK, in 1995. Currently, he serves as a member of the board of directors for the Software Engineering Society of Korea, the Computational Intelligence Society of Korea, and the Korean Society of Medical Informatics. He is currently a professor in the School of Computing at Korea Advanced Institute of Science and Technology (KAIST). His current research interests include artificial intelligence, data mining, software engineering, and biomedical informatics.