Available online at
www.mdl.mazedan.com

MCET

www.mazedan.com

www.mazedan.com/mcet

COMPARATIVE STUDY ON TEMPORAL INFORMATION EXTRACTION IN DIFFERENT LANGUAGES

PARUL PATEL

MAZEDAN COMPUTER ENGINEERING TRANSACTIONS
e-ISSN: 2583-0414
Article id: MCET0202003
Vol.-2, Issue-2
Received: 22 Sep 2021
Revised: 29 Oct 2021
Accepted: 10 Nov 2021

Citation: Patel, P. (2021). Comparative Study on Temporal Information Extraction in Different Languages. *Mazedan Computer Engineering Transactions*, 2(2), 7-11.

Abstract

In the world of internet, amount of digitized information is increasing exponentially. This data can be very used efficiently if meaningful information can be retrieved from it. Temporal information extraction is an emerging area of research in the field of natural language processing. Temporal information about document like document creation time, updation time, history of various updates on document in form of different versions can be very useful in the document clustering or timeline creation. Temporal information can be present in the form of temporal expression inside the document. It plays very important role in various applications like question answering, generating temporal summaries etc. Lot of research has been done in developing temporal information processing system, known as temporal tagger in various languages. In this paper, comparative study has been done on various temporal information processing system developed for different languages with their related issues and challenges.

Keywords: Temporal expression, Temporal tagger, TERN

1. INTRODUCTION

Information extraction is a process of extracting meaningful information from document which is generally in an unstructured form. Temporal information Extraction is the process of recognizing temporal cues from text that can be useful in applications like question answering, information retrieval, multi document summarization, timeline visualization.

Lot of work is done in clinical domain to forecast treatment effects patient timeline visualization, patients' selection for trials [1]. Due to the variety of type in which temporal information can be expressed in different language, extracting temporal expression becomes more challenging task. Temporal reasoning refers to a task of temporal expression recognition and interpretation of such expressions into a standard format for further use in different applications. Such task can also be referred as temporal tagger. This paper focuses on doing analysis of temporal tagger developed for different languages with their issues and challenges.

Approaches for Temporal Information Processing: Temporal information processing task is divided into two steps:

1. Extraction of temporal expression from raw text
2. Normalizing temporal expression into a standard value

Temporal information appears in raw text through temporal expression and event expression. Event expression is used to represent the events, while temporal expression is used to denote the time that can be classified into five different categories:

i. Explicit: It directly represents time (e.g., March 18, 2012, 14/11/2020)

ii. Implicit: It does not represent time value directly. It requires some reference time to be used while converting into standard value. Reference time can be either global or local. Local reference can be must be known to users for e.g., "Abdul Kalam died on 15th October 1931", and next statement "When he born, it was ....", Here 2nd sentence depends directly on earlier sentence. To interpret next statement, first statement is a reference point. When such references are available around text, it is known as local implicit expression. In global implicit reference, it is assumed users are aware of the reference or it must be added externally from knowledge base to normalize the temporal expression. For e.g., "Independence Day", "Diwali", "Onam" etc.

iii. Relative: Such expressions require to do calculation like add or subtract in to a given reference time. For e.g. "After two days"," two weeks ago"

iv. Vague: This type of expressions is ambiguous in nature and difficult to normalize. For e.g. ('early 1980's)

v. Non consuming reference: This type of expressions is not explicitly present in the document, but it is available in the form of metadata. For e.g., DCT (Document creation time), DMT (Document

Department of ICT, Veer Narmad South Gujarat University, Surat, Gujarat India
*Corresponding author email- pjpatel@vnsgu.ac.in

modification time), DAT (Document Access time) etc. This data plays very important role during normalization of temporal expressions. For e.g., news document contains its publication date. Temporal expressions can be present in above mentioned form in the text. It is necessary to convert it into standard format for further use. The structured forms of temporal information must convey the core information of the temporal expressions. A package of structure forms is called annotation language, because it is used to annotate the raw text [2]. In the context of temporal information extraction and interpretation, lot of research work is done in different languages. In this paper, following contribution is made:

- An introduction to annotation languages available for temporal information extraction
- An overview of different gold standard annotated dataset for TIE in different languages
- Comparison of various temporal taggers available for different languages

2. COMPARISON OF DIFFERENT ASPECTS OF TEMPORAL INFORMATION EXTRACTION

Annotation Languages

Temporal expression is usually present in the unstructured form. It is very important to convert it into standard template form for further use in applications like temporal information retrieval, question answering etc. In 2000, TIDES (Translingual Information Detection, Extraction, and Summarization) timex2 guidelines are given for temporal values as per ISO-8601[3]. In 2003, Time Markup Language (TimeML) was introduced that incorporated previous TIDES guidelines and Sheffield Temporal Annotation Guidelines (STAG) [4]. The ISO-TimeML was a standard version of TimeML, introduced in 2009 [5]. For Italian language, the Italian TimeML(It-TimeML) was proposed [6]. In TimeML and ISO TimeML, language diversity is not considered. For e.g., it was assumed that token level annotation is performed which is not applicable to different languages like Korean, Japanese, Chinese etc. Due to this reason, another Korean TimeML(K-TimeML) [7] was proposed in 2009 for korean language. Another revised version of K-TimeML was proposed to overcome the limitaitons of initial K-TimeML [8].

Dataset

TIDES published guideline for generating annotated dataset in 2001 to make dataset more consistent. It helps the researchers to evaluate their work with standard dataset. TempEval published various publicly available dataset including TimeBank. Following table summarizes various TIMEX or TimeML annotated gold standard dataset

Temporal Information Processing

Temporal information recognition and normalization (TERN) is also referred as temporal tagger in literature. It automatically extracts temporal expression from document or text and convert it into standard format. There are mainly three different approaches for temporal information extraction: (i) rule based (ii) machine learning based (iii) hybrid. In the following table, summary of various temporal tagger developed in different languages is listed with different approaches.

Table 1 List of available TIMEX annotated dataset

S.N	Dataset	Language	Description
1	TimeBank [9]	English	The TimeBank 1.2 Corpus contains 183 news TIMEX articles that have been annotated following the TimeML 1.2.1 specification. TimeBank 1.2 is free and is distributed by the Linguistic Data Consortium.
2	AQUAINT [10]	English	The AQUAINT TimeBank contains 73 news report TIMEX3 documents. It is very similar in content to, and uses the same specifications as, TimeBank 1.2
3	Wikiwar [11]	English	The Wikiwar corpus contains 22 annotated TIMEX2 documents collected from Wikipedia on topic world war.
4	TempEval [12]	English	There are 182 documents in a news report from television broadcasts, newswire or newspapers taken from several issues of the Wall Street Journal dating from 1989.
5	TimeBankPT [13]	Portuguese	Tempeval documents are translated into portuguese and then annotation is made.
6	Korean TimeBank [14]	Korean	Korean TimeBank consisting of more than 3,700 annotated sentences
7	KRAUTS	German	KRAUTS (Korpus of newspapeR Articles with Underlined Temporal expressionS) is a German corpus consists of two subsets: articles of the daily, regional newspaper DOLOMITEN and articles of the nationwide weekly newspaper DIE ZEIT. The corpus is composed of 192 documents with a total of 75,678 tokens [15].
8	Hindi TimeBank	Hindi	Hindi Timebank is a corpus of 1,000 articles with 25,829 events and 3,516 states for the purpose of temporal information retrieval in Hindi. The Hindi TimeBank has been created such that it can be used to further event annotation and detection research in Hindi, and the modifications to ISO-TimeML can be used to annotate TimeBanks for other Indo-Aryan languages. [16]
9	PersTimeML	Persian	It contains 43 documents from Peykareh and includes 26,949 tokens and 4,237 events. [17]
10	French TimeBank (FTiB)	French	The FTiB is a valuable resource that stimulate development and evaluation of French temporal processing systems, providing essential data for training machine learning systems. French TimeBank (FTiB), a corpus for French annotated in ISO-TimeML. It contains 109 documents with 16208 tokens. [18]
11	Spanish TimeBank 1.0	Spanish	It consists of 210 documents with over 75,800 tokens. It contains documents from AnCora Corpus [19]
12	ILTIMEX2012	Hindi	It consists of 300 manually tagged hindi news documents [20]

Table 2 Comparison of various temporal taggers available in different languages

S N	System Name	Language	TE recognition Approach	TE Normalization Approach	Description
1	HeidelTime	English, Spanish, French, Dutch, Italian, Chinese, Russian, German, Portuguese, Arabic	Rule Based	Rule Based	It is multilingual temporal tagger developed at Heidelburg University that extracts temporal expression from text and normalize it into TIMEX3 annotation standard.[21]
2	SUTime	English	Rule Based	Rule Based	Sutime is a temporal tagger developed by standford university. It is available as a library on various platforms [22]
3	TETI	Italian	Rule Based	-	It is a system that only recognizes temporal expressions [23]
4	Annotador	Spanish	Rule Based	Rule Based	It is Spanish temporal tagger. [24]
5	Portuguese System for Temporal Expression Recognition	Portuguese	Rule Based	-	It is a system designed to extract temporal expressions in a Portuguese language [25].
6	CTEMP	Chinese	Rule Based	Rule Based	The temporal parser that extracts and normalize comprehensive temporal expressions from Chinese texts. It is based on the chart parsing and constraint checking scheme [26]
7	CMedTex	Chinese	Rule Based	Rule Based	It is specifically designed for Chinese clinical document. [27]
8	ParsTime	Persian	Rule Based	Rule Based	ParsTime is a temporal tagger in Persian (Farsi) language. It is a rule-based system that extracts and normalizes Persian temporal expressions according to the TIMEX3 annotation standard [28]
9	INDTime	English	Rule Based	Rule Based	It is a temporal tagger designed to extract temporal expressions including festivals like Diwali, Holi and normalization by using exhaustive rule set [29]
10	ZamAn	Arabic	Rule Based	-	It is a rule based temporal extraction system developed for Arabic language.[30]
11	ATEL	English, Chinese	Machine Learning Based	-	A number of features are created from a predefined context centered at each token and augmented with decisions from a rule-based time expression tagger and/or a statistical time expression tagger trained on different type of text data, assuming they provide complementary information [31]

In this, different approaches for temporal information extraction and normalization are listed. Based on the comparative study, it seems that majority of the work is done either by using rule-based approach or by using data driven approach. Many attempts have been done for using machine learning and deep learning for this task. For extraction using machine learning approach requires features like window size of token, and POS of the token, n-grams of POS tag, whether a previous token is digit or not, previous token is temporal expression in a same window or not etc. It requires lots of linguistic observation of the specific language for feature engineering process. Moreover, language specific characteristics need to be utilized in a feature selection process.

3. CONCLUSION

Based on comparison, it is observed that lot of work has been done in the English language, but less work has been done in other languages especially regional, national languages. Other languages need more efforts due to unavailability of other resources like POS tagger, standard dataset etc. For e.g, to recognize temporal expression in regional language like Gujarati, there is a lack of POS tagger. In other regional languages, compared to English language resources, limited data sets are available for evaluation. In comparative study, it is observed that most of the systems are using rule-based approach in which handcrafted rules are applied to recognize temporal expressions. In future, more efforts can be made to improve performance and machine learning models can be designed with good features. Moreover, each language needs a temporal knowledge base (TKB) for interpretation purpose. Such TKB needs to be developed in future for resolving implicit temporal expressions like ‘last diwali’, ‘next onam’ etc.

REFERENCES

[1] Leeuwenberg, A., & Moens, M. (2020). A Survey on Temporal Reasoning for Temporal Information Extraction from Text (Extended Abstract). **IJCAI**.

[2] Chae-Gyun Lim, Young-Seob Jeong, & Ho-Jin Choi (2019). Survey of Temporal Information Extraction. **Journal of Information Processing Systems**, 15(4), 931-956. DOI: 10.3745/JIPS.04.0129.

[3] *Data elements and interchange formats – Information interchange –Representation of dates and times, ISO 8601, 2004*, ISO 8601, Data

elements and interchange formats – Information interchange –Representation of dates and times, 2004.

[4] A. Setzer, R. J. Gaizauskas, "Annotating events and temporal information in newswire texts," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, 2000; pp. 1287-1294.

[5] *Language resources management - Semantic annotation framework (SemAF) - Part1: Time and events, ISO 24617-1:2012, 2012*, ISO 24617-1, Language resources management - Semantic annotation framework (SemAF) - Part1: Time and events, 2012.

[6] T. Caselli, V. B. Lenzi, R. Sprugnoli, E. Pianta, I. Prodanof, "Annotating events, temporal expressions and relations in Italian: The It-TimeML experience for the Ita-TimeBank," in *Proceedings of the 5th Linguistic Annotation Workshop*, Portland, OR, 2011; pp. 143-151.

[7] S. Im, H. You, H. Jang, S. Nam, H. Shin, "KTimeML: specification of temporal and event expressions in Korean text," in *Proceedings of the 7th Workshop on Asian Language Resources*, Singapore, 2009; pp. 115-122.

[8] Y. S. Jeong, W. T. Joo, H. W. Do, C. G. Lim, K. S. Choi, H. J. Choi, "Korean TimeML and Korean TimeBank," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, Portoroz, Slovenia, 2016; pp. 356-359.

[9] Costa, Francisco and A. Branco. "TimeBankPT: A TimeML Annotated Corpus of Portuguese." **LREC** (2012).

[10] Lim, C., Jeong, Y., & Choi, H. (2018). Korean TimeBank Including Relative Temporal Information. **LREC**.

[11] Strötgen, J., Minard, A., Lange, L., Speranza, M., & Magnini, B. (2018). KRAUTS: A German Temporally Annotated News Corpus. **LREC**.

[12] Goel, P., Prabhu, S., Debnath, A., Modi, P., & Shrivastava, M. (2020). Hindi TimeBank: An ISO-TimeML Annotated Reference Corpus. **ACL 2020**.

[13] Yaghoobzadeh, Y., Ghassem-Sani, G., Mirroshandel, S.A., & Torbati, M. (2012). ISO-TimeML Event Extraction in Persian Text **COLING**.

[14] André Bittar, Pascal Amsili, Pascal Denis, Laurence Danlos. French TimeBank: An ISO-TimeML Annotated Reference Corpus. ACL 2011 - 49th Annual Meeting of the Association for Computational Linguistics, Jun 2011, Portland, Oregon, United States. pp.130-134.

[15] Saurí, Roser & Badia, Toni. (2012). Spanish TimeBank 1.0.

[16] Ramrakhiyani N., Majumder P. (2013) Temporal Expression Recognition in Hindi. In: Prasath R., Kathirvalavakumar T. (eds) Mining Intelligence and Knowledge Exploration. Lecture Notes in Computer Science, vol 8284. Springer, Cham. https://doi.org/10.1007/978-3-319-03844-5_72

[17] Strötgen, Jannik & Gertz, Michael. (2015). A Baseline Temporal Tagger for all Languages. 541-547. 10.18653/v1/D15-1063.

[18] Angel X. Chang and Christopher D. Manning. 2012. SUTIME: A Library for Recognizing and Normalizing Time Expressions. *8th International Conference on Language Resources and Evaluation (LREC 2012)*.

[19] F. dell'Orletta, and I. Prodanof T. Caselli, "TETI: a TimeML compliant TimEx tagger for Italian" in Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT), Mragowo, Poland, 2009-October

[20] Navas-Loro, María and Rodríguez-Doncel, Víctor. 'Annotador: a Temporal Tagger for Spanish'. 1 Jan. 2020: 1979 – 1991.

[21] C. Hagège, J. Baptista and N. Mamede, "Portuguese Temporal Expressions Recognition: From TE Characterization to an Effective TER Module Implementation," 2009 Seventh Brazilian Symposium in Information and Human Language Technology, 2009, pp. 36-43, doi: 10.1109/STIL.2009.12.

[22] Wu, Mingli & Li, Wenjie & Lu, Qin & Li, Baoli. (2005). CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information. 3651. 694-706. 10.1007/11562214_61.

[23] Liu, Z., Tang, B., Wang, X., Chen, Q., Li, H., Bu, J., Jiang, J., Deng, Q., & Zhu, S. (2017). CMedTEX: A Rule-based Temporal Expression Extraction and Normalization System for Chinese Clinical Notes. *AMIAAnnual Symposium proceedings. AMIA Symposium*, *2016*, 818–826.

[24] Mansouri, Behrooz & Zahedi, Mohammad & Campos, Ricardo & Farhoodi, Mojgan & Rahgozar, Maseud. (2018). ParsTime: Rule-Based Extraction and Normalization of Persian Temporal Expressions. 10.1007/978-3-319-76941-7_67.

[25] Patel, Parul & Patel, S. (2016). INDTime: Temporal Tagger—First Step Toward Temporal Information Retrieval. 10.1007/978-981-10-0129-1_21.

[26] Saleh I., Tounsi L., van Genabith J. (2011) ZamAn and Raqm: Extracting Temporal and Numerical Expressions in Arabic. In: Salem M.V.M., Shaalan K., Oroumchian F., Shakery A., Khelalfa H. (eds) Information Retrieval Technology. AIRS 2011. Lecture Notes in Computer Science, vol 7097. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-25631-8_51

[27] Y. Chen, and B. Douglas K. Hacioglu, "Automatic time expression labeling for English and Chinese text" in Pro. of the 6th Int. Con, on Intelligent Text Processing and Computational Linguistics (CICLing), vol. 3406 of LNCS, Mexico City, Mexico., February 2005, pp. 548-559.