Inference of Absolute Time Value from Temporal Expressions

Junehwan SUNG
Individual
Tokyo, Japan
https://orcid.org/0000-0002-1668-4377

Shinsuke MORI
Academic Center for Computing and Media Studies,
Kyoto University
Kyoto, Japan
https://orcid.org/0000-0001-8596-8667

Hirotaka KAMEKO
Academic Center for Computing and Media Studies,
Kyoto University
Kyoto, Japan
https://orcid.org/0000-0001-9844-6198

Akira KUBO
Graduate School of Arts and Sciences,
The Open University of Japan
Chiba, Japan
https://orcid.org/0000-0002-4901-8290

Tatsuki SEKINO
The International Research Center for Japanese Studies,
National Institutes for the Humanities
Kyoto, Japan
sekino@nichibun.ac.jp

***Abstract*—In this paper, we explore and discuss a way to extract temporal information from natural language texts. The suggested method is divided into two parts: temporal expression recognition and temporal value inference. The former employs the conventional NER approach, using a BiLSTM-CRF architecture. The latter is implemented with a rule-based algorithm, which can be further developed in later work for better coverage of various temporal expressions. In terms of the corpus, we have selected 200 articles from one of the major Japanese newspaper companies to create an annotated corpus, classifying temporal expressions into five different types. As for the performance, we have achieved 0.866 in F-measure for the recognition of temporal expressions and 0.920 in accuracy for the inference of the absolute temporal values of the expressions. Combining the two modules and running them as an end-to-end system, we have attained 0.891 of F-measure.**

***Index Terms*—temporal expression, absolute time value, named entity recognition**

I. Introduction

The amount of information various types of natural language texts contain is immeasurable, and along with the development of digitizing and storing technology, such data have become more and more accessible. However, due to the complexity of analyzing them, we have not been able to fully appreciate the vastness of data and the information we could extract from them.

One of the factors that renders analyzing natural language texts difficult is that it is hard to comprehend relations among data. When becoming able to resolve temporal expressions and to specify the absolute time point they are referring to, we can visualize texts on a time line [1] as shown in Fig. 1 and infer temporal relations among different data using Allen's interval algebra [2]. These visualizations and inferences allow us to discover correlations of data.

In this paper, as part of exploring ways to exploit the benefit of big data, we present

1) a temporal corpus in which temporal expressions (word sequences) are annotated manually with absolute time,
2) a method based on a machine learning technique for recognizing temporal expressions in texts, and
3) a rule-based method for inferring the absolute time of temporal expressions.

We developed these methods, tested on our corpus, and confirmed their validity.

In the following sections, we are going to start with introducing related works. Then, we describe our corpus along with the annotation standard we used, followed by explanations on our methods employed for recognizing temporal expressions and inferring their absolute temporal values. After evaluating the proposed methods using our corpus, we analyse the results. Finally, we conclude the paper with the summary and propose possible future work.

II. Related Work

As we have mentioned, our contributions can be divided into three parts: an annotated corpus, a time expression recognizer, and an absolute time inferring system. In this section, we introduce some previous studies related to these contributions.

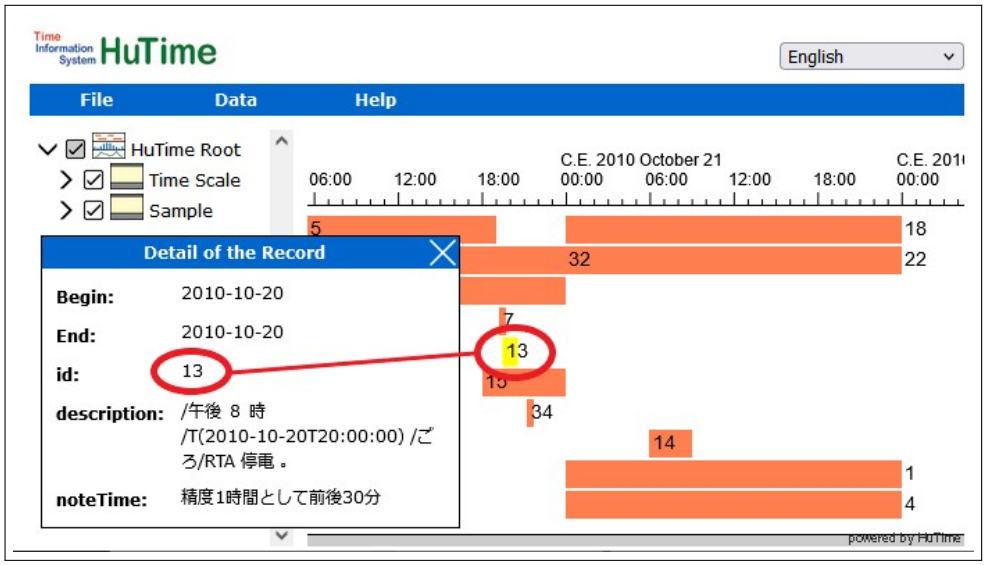

Fig. 1. Example of temporal expression visualization by HuTime [1].

A. Annotated Corpus

One of the early attempts of corpus construction for natural language processing (NLP) is the Penn Treebank [3], in which each word is annotated with a part-of-speech, and each sentence is annotated with a syntactic structure. Since our target – time point expressions – is a sequence of words, a corpus containing annotated named entities (NEs) [4] is more appropriate for our purpose. The set of NEs includes, along with person names, organization names, time expressions, and date expressions. The time expressions and date expressions in this study correspond to time points or duration. In our paper, we concentrate exclusively on expressions corresponding to a time point. In addition, we do not distinguish time expressions and date expressions from each other, as both of them refer to a time point with different specification levels.

Our corpus has the absolute time value for temporal expressions, and the temporal expressions are classified into different types, which are explained later in this paper. It is similar to the corpus from SemEval-2010 Task 13 [5], where some expressions have a `VAL` attribute according to an extension of the ISO 8601 standard. To be more accurate, our annotation takes into account the observation that an arbitrary temporal expression referring to a time point corresponds to a span on the time axis, rather than pinpointing a specific time point. For example, by the expression "at 13:00 on the 18th of January," the author refers to a time point lying *from and including* the start point of 13:00 on that day *to but excluding* the start point of 13:01 on the same day.

B. Temporal Expression Recognition

In SUTime [6], another previous study for resolving temporal expressions, a rule-based method has been proposed. SUTime explicitly mentioned, as one of the limitations, its incapability of dealing with ambiguous phrases, which results in capturing "spring" as in *"The water from the spring was fresh and clear"* due to its rule-based methodology.

As a solution to the limitation, we propose a machine learning technique which resolves various ambiguities, referring to the context. By our definition, a temporal expression is a sequence of words, which makes recognizing them almost identical to the named entity recognition (NER) in NLP. NER has a long research history [4] for general NEs. In recent days, the framework has been widely applied to bio-medical texts [7] and even cooking recipes [8], among others.

Various solutions have been proposed for NER such as maximum entropy-based [9], support vector machine-based [10], logistic regression-based [11], and Conditional Random Fields (CRF)-based [12]. CRF-based approaches show high performances and are often used for NER. These days, deep neural network-based approaches, such as Bidirectional Long Short-Term Memory (BiLSTM), are focused on feature extraction of sentences.

In this paper, we adopt BiLSTM-CRF, a combination of BiLSTM and CRF, which can recognize the meaning of words based on their contexts. As a result, our model can differentiate 震災25 年 (*en:* 25 years later since the earthquake) and 25 年 (*en:* the year of 2025) from each other.

C. Absolute Temporal Value Inference

SUTime [6] also infers the absolute value of temporal expressions. Our system has a similar module to infer it, limiting the scope to specific time spans expressed by a left substring of *YYYY-MM-DD*T*hh:mm:ss*. Unlike SUTime, we do not employ other expressions, such as "WE" for weekends, because these expressions require some more disambiguation to enable visualization [1] (see Fig. 1) or inference based on Allen's interval algebra [2]. We leave the absolute time inference for these expressions, such as "winter" and "morning", as our future work.

Lastly, our research embraces novelty value in the aspect that it supports the Japanese language, and this is the first work done in Japanese. There has been an annotation work

done [13], but no attempts of recognition and inference that we know of. We should, however, emphasize that the support of the Japanese language does not necessarily limit the expandability of our system to another language.

III. Data Preparation

To develop the absolute value estimator for temporal expressions and evaluate its performance, we have created our own annotated corpus. In this section, we describe the list of the tags we used to annotate temporal expressions. Then, we explain the definition of the absolute temporal value assigned to each temporal expression. Finally, we share the statistics of our corpus.

A. Type of Temporal Tags

As we described, our goal is to develop a method for estimating the absolute values for temporal expressions in texts. We have observed various expressions conveying a time-related meaning and classified them into five types, as listed in TABLE I: one (T) for absolute expressions describing a time point directly, and four (RTI, RTF, RTP, and RTA) for relative expressions describing a time point based on another time point. We excluded expressions indicating duration, as done in TempEval-2 task (SemEval-2010 Task 13) [5]. In the following section, we explain these tags one by one.

1) T*: Absolute Time:* The most straightforward way to describe a time point is to specify it explicitly. The specification level varies, and we regard the whole sequence of words as one entity as long as they are referring to one temporal expression. Here is an example:

Example (en)

The deadline of paper submission is /31st of October 2021/T.

Example (ja)

/平成 23 年 3 月 11 日 14 時 46 分 18.1 秒/T に、日本の三陸沖の太平洋を震源とする巨大地震が発生した。

Sometimes, the year or other higher temporal concepts are omitted as in the following example.

Example (en)

The 2011 Tōhoku earthquake and tsunami occurred at /14:46 on 11th of March/T.

In this English example, the year is obvious from the context, and the author omitted it even though s/he must have had it in mind.

2) RTI *(Relative Time Internal):* The tag RTI represents relative temporal expressions indicating an internal span of a certain *time point*, which corresponds to another span. For example,

Example (en)

In /the morning/RTI of the 2011 Tōhoku earthquake day, there were small earthquakes ...

indicates a certain time span completely included in the day of March 11th in 2011, which is implicitly specified.

3) RTF *(Relative Time Future):* represents relative time which indicates the future with regard to the reference time point e.g.,

Example (en)

/four days later/RTF ...

4) RTP *(Relative Time Past):* represents relative time which refers to the past in regard to the reference time point e.g.,

Example (en)

/two years ago/RTP ...

5) RTA *(Relative Time Around):* mostly assigned to words that indicate the proximity to the time they are modifying e.g.,

Example (en)

/Around/RTA /3 p.m./T ...

B. Notation of Absolute Temporal Value

As for the annotation convention for describing temporal values, we have adopted the ISO 8601[1] standard as the base notation, as it is widely known and compatible with various libraries i.e., easily convertible to different data types or formats, using Python, for instance. As a result, you can find the following types of annotations in our corpus:

1) (2021-09-08T00:00:00)
2) (2021-05-05T00:00:00,2021-05-08T00:00:00).

1) represents a specific time point, 8th of September 2021 in this case, and 2) indicates a time span that starts from 5th of May to the end of 7th of May 2021. For succinctness, all time units except the year can be omitted, when not specified in the expression e.g., (2021-09-08) for "8th of September 2021." However, what we refer by "2021-09-08" in this paper should be interpreted as a span, i.e., [2021-09-08T00-00-00.00..., 2011-09-09T00-00-00.00...).

Another point we would like to clarify is the difference between the following two notations:

1) [2021-09-08T00:00:00, 2021-09-08T00:00:01)
2) [2021-09-08T00:00:00, 2021-09-08T00:00:02)

As for 1), we annotate as (2021-09-08T00:00:00), while 2) as (2021-09-08T00:00:00, 2021-09-08T00:00:02).

Lastly, although ISO8601 explicitly defines the time zone, we have decided to leave out the time zone notation as well, considering the nature of the corpus originally targeting at Japanese readers. Therefore, all temporal values shall be considered to be of Japan Standard Time (UTC+09:00).

[1]https://www.iso.org/iso-8601-date-and-time-format.html

Type	Tag	Description	Example
Absolute	T	a specific time point (basis time point)	2015/Mar/15, today 3 月 15 日 に ...
Relative	RTI	a relative expression indicating a span included by the basis time point (span)	In the morning on the 4th ... 4 日 の 朝 ...
	RTF	a relative expression indicating the move to the future direction from the basis time point	And three days later ... その 3 日 後 ...
	RTP	a relative expression indicating the move to the past direction from a time point	Five years ago ... 5 年 前 ...
	RTA	a relative expression indicating a span near to the basis point	At around three o'clock ... 3 時 頃 ...

TABLE I
TAG LIST.

	#Articles	#Sent.	#Tokens	#TEs	(T	RTI	RTF	RTP	RTA)
Earthquake	100	1308	39821	439	242	20	140	34	3
Flood	100	1409	43970	766	517	152	29	25	43
Total	200	2717	83791	1205	759	172	169	59	46

TABLE II
SPECIFICATIONS OF OUR TIME-ANNOTATED CORPUS.

C. Annotated Corpus

We selected and annotated 200 Japanese news articles from a newspaper (The Mainichi Shimbun[2]) dated from 2010 to 2019. The sentences in the articles are tokenized using KyTea [14], [15] prior to the annotation. For the detailed statistics of the corpus, please refer to TABLE II.

We selected 100 articles, each of which is related to earthquake and flood disasters, by using Latent Dirichlet Allocation (LDA) [16]. First, using nouns, verbs, adverbs, adjectives, and adjectival nouns, we trained LDA models, and subsequently selected most frequent 100,000 words which are used only in less than 50% articles. We set α as

$$\alpha_i = \frac{1.0}{i + \sqrt{N}}, \tag{1}$$

where i shows the index of the topic and N shows the number of topics. We used a model with $N = 256$.

Then, we chose two topics manually; one is related to earthquake, and the other is related to flood damage. By using the model, we calculated the distributions of topics of each article. Finally, we selected 100 articles with highest weights of each topic and of the size greater than 4,096 bytes.

IV. ABSOLUTE VALUE INFERENCE FOR TEMPORAL EXPRESSIONS

We suggest a method to infer the absolute time values for temporal expressions within texts, which can be further segmented into two parts: temporal expression recognition and absolute temporal value inference. The recognition is trainable regardless of languages, while the absolute value inference is tuned to the Japanese language. A detailed explanation of each module will follow the overview of the whole process described below.

1) Input: The system takes a tokenized sentence as the input.

Example (en)

The 2011 Tōhoku earthquake and tsunami occurred at 14:46 on 11th of March.

Example (ja)

平成 23 年 3 月 11 日 14 時 46 分 18.1 秒 、 三陸 沖 を 震源 と する 大 地震 が 発生 し た 。

2) Process 1: The temporal expression recognition module recognizes word sequences expressing a time point.

Example (en)

The 2011 Tōhoku earthquake and tsunami occurred at /14:46 on 11th of March/T.

Example (ja)

/平成 23 年 3 月 11 日 14 時 46 分 18.1 秒/T 、 三陸 沖 を 震源 と する 大 地震 が 発生 し た 。

3) Process 2: The absolute value inference module estimates the absolute value for each time point expression.

Example (en)

/14:46 on 11th of March/T
→ 2011-03-11T14:46:00

Example (ja)

/平成 23 年 3 月 11 日 14 時 46 分 18.1 秒/T
→ 2011-03-11T14:46:18.1

4) Result: The system returns the sentence with recognized temporal expressions each of which is annotated with an absolute value.

[2]https://mainichi.jp/english/ (accessed on October 24th, 2021).

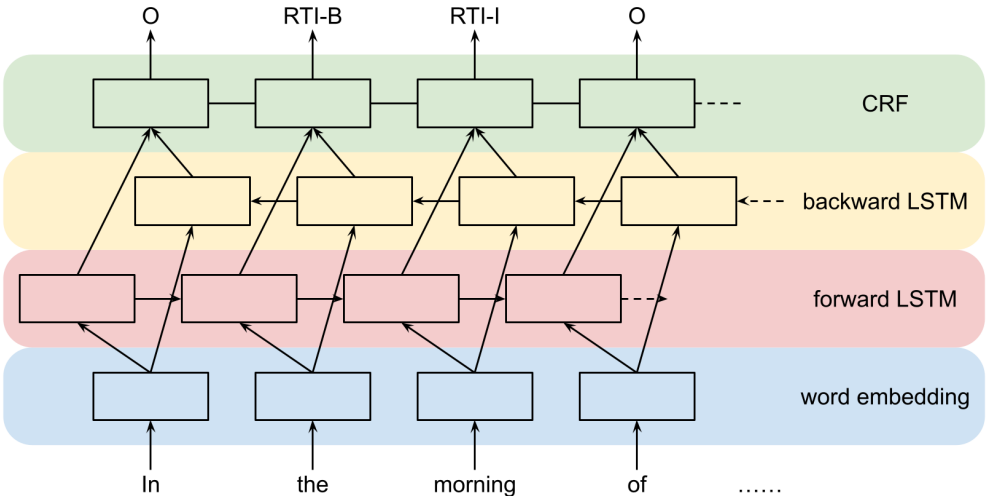

Fig. 2. Overview of a BiLSTM-CRF model.

Example (en)

The 2011 Tōhoku earthquake and tsunami occurred at /14:46 on 11 March/T(2011-03-11T14:46:00).

Example (ja)

/平成 23 年 3 月 11 日 14 時 46 分 18.1 秒/T(2011-03-11T14:46:18.1) 、 三陸 沖 を 震源 と する 大地震 が 発生 し た 。

In the subsequent parts to this section, we are going to explain these two processes in detail.

A. Temporal Expression Recognition

As mentioned above, Process 1 – temporal expression recognition – is almost identical to the conventional named entity recognition, or NER. Similar to NER solutions we adopt word-based sequence labelling framework, in which the first token of an expression (word sequence) of the type X is labelled with X-B, X-I for the following ones, and O for all other tokens (see Fig. 2).

For the sequence labelling, we construct BiLSTM-CRF models for NER. Fig. 2 shows the overview of the model. As its name suggests, BiLSTM-CRF is a combination of BiLSTM and CRF. LSTM [17] is a variant of recurrent neural network architectures, and is capable of capturing long-distanced dependencies. BiLSTM is a type of architecture which consists of a forward LSTM and a backward LSTM. CRF [18] is a model often used for structured prediction. With a CRF layer, we can treat sequential dependencies of predictions.

With BiLSTM-CRF, we first embed each word into vector space. Then, we encode them by using BiLSTM. Lastly, we calculate the probabilities of sequences of labels with the CRF layer, and output the best sequence of labels.

B. Absolute Value Inference

Based on the results of the named entity recognition done above, we perform an inference of absolute time values. In this paper, as a starting point, we have decided to narrow the inference scope to the temporal expressions tagged as T only, which is the most frequent and most important. The other tags are going to be supported in later work.

1) Document creation time (DCT) [6] is a prerequisite of inference.
2) Filter non-number temporal expressions e.g., today, of which time point can be specified without the following rules.
3) If the expression provides necessary figures for each time unit and date, use them.
4) As for the expressions that lack higher-level time concepts – just “March,” instead of “March 2001,” we supplement what is missing from the previous temporal expression. In other words, having n number of T-tagged temporal expressions and notating each of the expressions as $e_1, e_2, \cdots, e_t, \cdots, e_n$, we will refer to e_{t-1} to complement e_t. For e_1, DCT will be its reference point, or e_0.
5) When only the day is specified e.g., 24th, compare the following three possible time points and choose the nearest one from the date when the article was issued:
   a) the same day of the previous month
   b) the same day on the month of DCT
   c) the same day on the month of e_{n-1}
6) Following the annotation convention discussed in Section III-B, we leave out all the lower time units than what is specified e.g.,“11th of March” has day as its lowest, or most granular time unit within the given expression, which allow us to omit hour, minute, and second.

V. Evaluation

In order to evaluate our method, we conducted the following experiments on our corpus.

Type	Precision	Recall	F-measure
T	0.913	0.928	0.920
RTI	0.865	0.809	0.830
RTF	0.785	0.684	0.729
RTP	0.760	0.628	0.677
RTA	0.957	0.798	0.866
All (weighted)	0.883	0.856	0.866

TABLE III
TEMPORAL EXPRESSION RECOGNITION MEAN ACCURACY OF FIVE FOLDS.

Dimension of word embedding	300
Dimension of LSTM hidden state	256 × 2 (forward and backward)
Batch size	30
Number of epochs	300

TABLE IV
HYPER-PARAMETERS FOR TRAINING OUR NER.

- Temporal expression recognition for tokenized sentences (Process 1).
- Absolute time inference on correctly recognized temporal expressions, which are manually picked out (Process 2).
- An end-to-end evaluation from the recognition of temporal expressions to the inference of absolute temporal values for the recognized expressions (Overall).

Since Process 1 is realized by a neural network approach that requires training data, we have divided the corpus into a training set and a test set, and have applied the k-fold cross-validation technique with $k = 5$.

On the contrary, Process 2 does not require training, and we evaluated all the T-tagged temporal expressions in the corpus. To reiterate, we concentrate the scope of this work on T-tagged expressions only. The evaluation of the overall process, or the end-to-end process, is also limited to T-tagged expressions.

A. Process 1

First, we trained a NER model of the BiLSTM-CRF architecture. The hyper-paramters used for training are shown in TABLE IV.

To evaluate the performance of the temporal expression recognition module, we used the same evaluation metrics as those for general NER models, namely recall, precision and F-measure, as described in TABLE III. Additionally, the confusion matrix of the tags at IOB-level is also mapped in Fig. 3.

First, we can observe that the F-measures are high enough, considering the F-measure of NER models in the general domain tend to be around 0.9 with the training data of about 10,000 sentences (our case: 4/5 × 2,717 sentences). From TABLE III and Fig. 3, we can derive that F-measures of RTF and RTP are relatively low compared to the others, and they are often misclassified as O and T, respectively. Having looked at the actual data, we learned that RTF and RTP have more literal variations, such as 一夜 明け (*en*: *lit.* break of one night; next day), while the other tags have set patterns of literals e.g., numerals followed by time units would most likely be T, if not always. Therefore, we presume the errors are going to be addressed with more data.

B. Process 2

For the evaluation of the absolute temporal value inference (Process 2), we have extracted T-tagged temporal expressions from the whole corpus, which sum to 759.

Using the four principles described in Section IV-B, we have performed the inference and achieved the accuracy of 0.920 (698/759). Analysing the missed cases, we have discovered a few patterns that our system cannot properly deal with:

1) The logic is likely to fail when the author jumps between the past and the future without explicitly mentioning the alteration of the time frame. This is due to the fact that the inference logic of missing time units is based on the previous one of the target temporal expression.

> Article date: 2019-03-01
>
> **[Previous T]** In the afflicted area of the Great Hanshin earthquake occurred in /1995/T(1995) ...
> **[Target T]** The exhibition lasts until /30th of April/T(2019-04-30).

In the above example, the system returned "1995-04-30" for the target expression, as it referred to the previous expression for its missing year, whereas the answer should be "2019-04-30".

2) The system cannot differentiate direct speeches described in the article from the main content.

> Article date: 2011-09-05
>
> (INTERVIEW): "... I have never had to do the actual evacuation before this, but it seems like I still need to go through another day here /today/T(2011-09-4)."

The time point that the word "today" can refer to varies depending on the speaker or the context. However, we cannot enumerate all the possible patterns, and "today" always takes the article date as its temporal value, which could result in the wrong inference as shown here.

3) The reference point is marked with another tag other than T.

> Article date: 2014-11-23
>
> **[Previous T]** The Great East Japan earthquake occurred on /March 2011/T(2011-03) ...
> **[Previous temporal expression]** The exhibition will display /20 years/RTF(2015) from the Great Hanshin earthquake...
> **[Target T]** The action plans of the 20th memorial day celebration will be revealed on /18th of January/T(2015-01-18) ...

The system inferred the temporal value of the target expression as "2011-01-18", referring to the previous T. However, the author changed the context by mentioning the upcoming year would be 20th year since the Great Hanshin earthquake, which is the actual time point that the target expression is referring to. Since we have

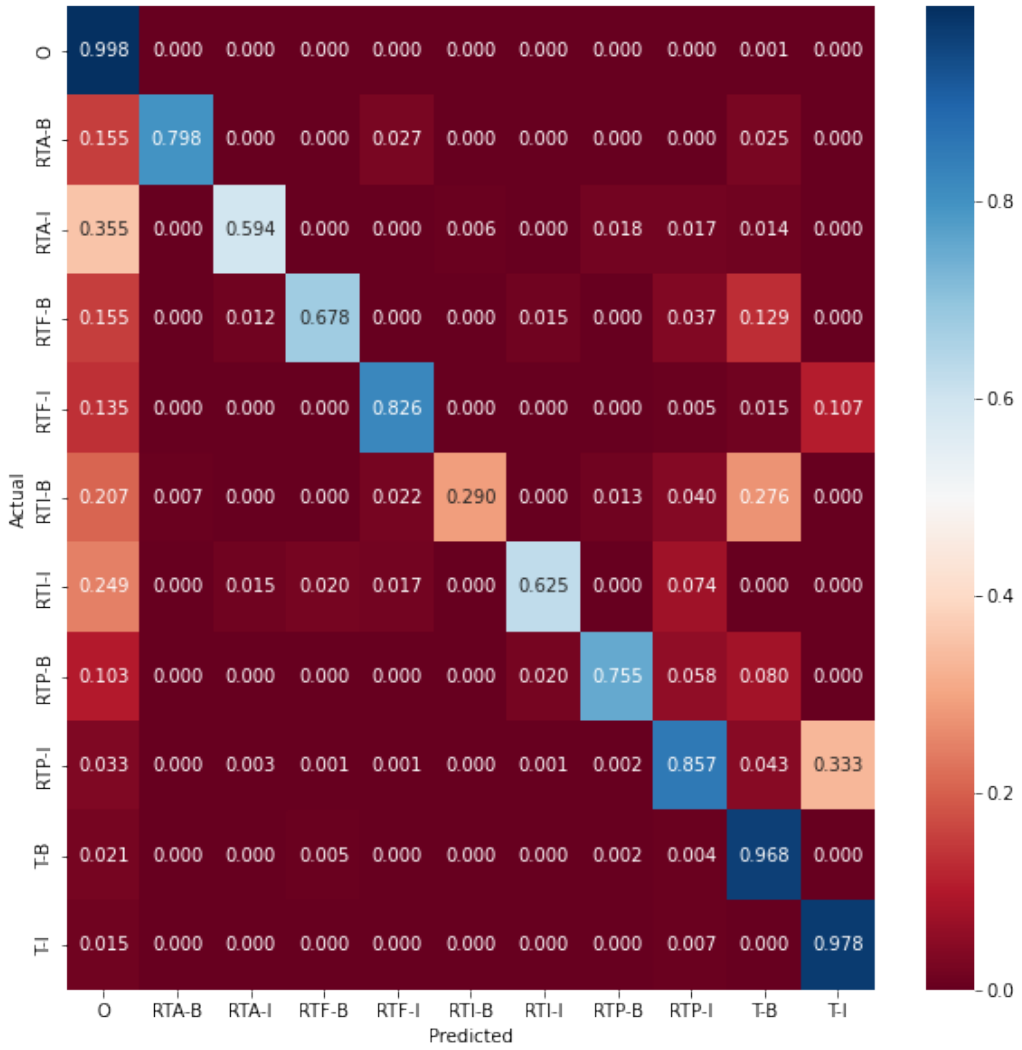

Fig. 3. Mean IOB accuracy of five folds

limited the types of the tags to T this time, the system could not refer to the right temporal expression.

On top of these main patterns, other missed cases include 当日 *(en: that day)* or 前日 *(en: previous day)*, which cannot be determined by rules. We expect the performance to increase with probabilistic models that can incorporate contextual information or syntactic information, such as grammatical relations among words.

C. Overall for T

For an overall evaluation, we inferred the absolute time for T-tagged expressions recognized and classified in Process 1, and calculated the accuracy. As a result, we have acquired 0.888, 0.894, and 0.891 for the precision, recall, and F-measure, respectively.

This F-measure is close to the multiplication of the F-measure of T of Process 1 and the accuracy of Process 2. Thus, we can safely presume that improvements of the both modules contribute to overall improvement.

In addition, the overall accuracy is sufficient for us to try, in parallel with the improvement, visualization and inference of temporal information for a text archive.

VI. Conclusions

As one way of benefiting from a considerable amount of data, we could establish temporal relations among them. Even though all data have creation dates, not all the dates of the events described within are necessarily clear enough to be specified in a scalable manner. To exploit and analyze big data, we have proposed a method that can recognize temporal expressions within natural language texts and extract temporal information included. Our proposed method consists of two main phases: recognition and inference.

Within the recognition phase, we use a BiLSTM-CRF architecture to recognize temporal expressions and assign appropriate tags from the following five types: T, RTI, RTF, RTP, and RTA. To train the network, we have annotated 200 articles from a Japanese newspaper and used 160 articles for

training; the remaining 40 articles were used for testing. As a result, our system achieved 0.886 of F-measure, which is the mean results of the k-fold cross validation with $k = 5$.

As for the inference phase, we implemented a rule-based system that can extract absolute temporal values from temporal expressions. Although the system is a combination of a few simple rules, we have attained 0.920 of accuracy, when tested on 200 newspaper articles, or 759 T-tagged temporal expressions. Although we have focused only on T-tagged expressions in this paper, we will develop our work to support the remaining tags in the following research.

Having analyzed the errors from each process, we have learned that the following actions can improve the performance. As for the recognition module, more data which will provide more literal variations for each tag can boost its performance. When it comes to the inference module, incorporating contextual or grammatical information would be useful, in addition to the current mutual reference among expressions and the use of document dates. We would like to attempt these in the follow-up studies.

Through this paper, we have demonstrated our ways of recognizing temporal expressions and inferring the absolute temporal values to which the expressions are referring. Extracting such information opens up the accessibility to big archival data and facilitate their analyses, allowing us to establish temporal relationships across different events expressed in collections of records. With the time information extracted, we can locate events described in different sources on the same timeline and sort them in the order of occurrence. We could also extract all the events concerning a certain time point, or analyze easily the time period a certain collection covers, for instance.

These days, using archival or repository systems, we can find and access most of data we look for, but it is not always easy to extract only necessary or relevant information from them, and it often requires a high extent of manual labor, especially when you are not familiar with the text. By enabling intrinsic temporal information, we expect our approach to contribute to increasing the interpretability and the usability of huge collections of written records.

REFERENCES

[1] T. Sekino, "Time information system, HuTime - A Visualization and Analysis Tool for Chronological Information of Humanities," in *Proceedings of Digital Humanities Conference 2020*, 2020.

[2] J. Allen, *Communications of the ACM*, vol. 26, no. 11, p. 832–843, 1983.

[3] M. Marcus, G. Kim, and M. A. Marcinkiewicz, "The Penn Treebank: A Revised Corpus Design for Extracting Predicate Argument Structure," 1993, pp. 77–81.

[4] N. A. Chinchor, "Overview of MUC-7/MET-2," 1998.

[5] M. Verhagen, R. Saurí, T. Caselli, and J. Pustejovsky, "SemEval-2010 task 13: TempEval-2," in *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 57–62. [Online]. Available: https://aclanthology.org/S10-1010

[6] A. X. Chang and C. Manning, "SUTime: A library for recognizing and normalizing time expressions," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 3735–3740. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/284_Paper.pdf

[7] T. Erjavec, J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Encoding biomedical resources in TEI: The case of the GENIA corpus," in *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*. Sapporo, Japan: Association for Computational Linguistics, Jul. 2003, pp. 97–104. [Online]. Available: https://aclanthology.org/W03-1313

[8] S. Mori, H. Maeta, Y. Yamakata, and T. Sasada, "Flow graph corpus from recipe texts," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 2370–2377. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/763_Paper.pdf

[9] A. E. Borthwick, "A Maximum Entropy Approach to Named Entity Recognition," Ph.D. dissertation, USA, 1999, AAI9945252.

[10] M. Asahara and Y. Matsumoto, "Japanese Named Entity Extraction with Redundant Morphological Analysis," in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, pp. 8–15. [Online]. Available: https://aclanthology.org/N03-1002

[11] T. Sasada, S. Mori, T. Kawahara, and Y. Yamakata, "Named Entity Recognizer Trainable from Partially Annotated Data," in *Conference of the Pacific Association for Computational Linguistics*. Springer, 2015, pp. 148–160.

[12] A. McCallum and W. Li, "Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 188–191. [Online]. Available: https://aclanthology.org/W03-0430

[13] M. Asahara, S. Yasuda, H. Konishi, M. Imada, and K. Maekawa, "BCCWJ-TimeBank: Temporal and Event Information Annotation on Japanese Text," in *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*. Taipei, Taiwan: Department of English, National Chengchi University, Nov. 2013, pp. 206–214. [Online]. Available: https://aclanthology.org/Y13-1019

[14] G. Neubig and S. Mori, "Word-based Partial Annotation for Efficient Corpus Construction," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/408_Paper.pdf

[15] G. Neubig, Y. Nakata, and S. Mori, "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 529–533. [Online]. Available: https://aclanthology.org/P11-2093

[16] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," vol. 3, 01 2001, pp. 601–608.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the 18th International Conference on Machine Learning*, 2001, pp. 282–289.

VII. ACKNOWLEDGMENTS

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 20H00017 and 20H04210. We are also grateful to the annotators for their contribution to the design of the guidelines and to the annotation.