TimeInfo: a Semantic Annotation Framework for Temporal Information in Scientific Papers

Salah Yahiaoui, Iana Atanassova

TimeInfo: a Semantic Annotation Framework for Temporal Information in Scientific Papers

Salah Yahiaoui [0000−0001−5483−1764]
and Iana Atanassova [0000−0003−3571−4006]

CRIT, Université de Bourgogne Franche-Comté
30 rue Megevand 25000 Besançon, France
salah.yahiaoui@edu.univ-fcomte.fr, iana.atanassova@univ-fcomte.fr

Abstract. In this article we propose a new schema for the annotation of temporal expressions in texts that we name TimeInfo. Also, we present and analyse existing models for temporal data annotation. Then, we define the set of elements and their attributes in TimeInfo and explain how these elements provide finer distinctions for the annotation of temporal expressions. The main purpose of our model is to allow for better detection, extraction, and semantic representation of temporal data in scientific papers. Finally, we present the use cases and the different applications that can be developed using the TimeInfo framework.

Keywords: TimeInfo · Temporal information · Semantic annotation · Mining scientific papers · Text mining · Information extraction · TimeML.

1. Introduction

Today, the availability of open access textual data and the growing computing power allow for the processing of large corpora. In our research, we focus on scientific papers and more precisely on the pro-

cess of annotating temporal data in scientific papers. In this paper, we discuss existing temporal data annotation systems such as TIMEX2 and TIMEX3. Then, we introduce our own annotation framework, called TimeInfo, that has been designed to provide a richer semantic representation of temporal information extracted from scientific papers. We present our annotation scheme, discuss its differences with the existing schemes, and propose a first methodology for the annotation with TimeInfo.

The syntax of TimeInfo is accessible and flexible, which allows for various applications at a large-scale. TimeInfo can be used to develop tools for the extraction and semantic annotation of temporal data, temporal data visualization and the improvement of search engines. TimeInfo's core is based on the semantic value of the temporal information and its co-text. While TimeInfo is initially designed for the English language, it can be adapted to other languages, since the value of temporal information does not change from one language to another.

2. Overview of existing annotation schemes

The task of temporal data recognition is a subcategory of Named Entity Recognition.

Historically, one of the first temporal data annotation systems was introduced in the MUC-7 conference which is sponsored by the Defense Advanced Research Projects Agency (DARPA) [Chinchor, 1998]. Thus, MUC-7 introduced the first standard of temporal data recognition and annotation using the XML TIMEX tag. This first standard provides a very simple representation of expressions, as the TIMEX tag has only one attribute that is TYPE and two possible values that are DATE and TIME. Then, in 2003, the TIMEX2 annotation system was born thanks to the TIDES research program [Ferro *et al*., 2003]. TIMEX2 was the first major upgrade to TIMEX, with the main purpose to provide a set of guidelines for constructing a temporal data annotation scheme.

Later, the TimeML annotation scheme has been developed in the TERQAS workshop [Saurí *et al*., 2006] for the annotation of events, temporal expressions, and the links that they share. Further, TimeML

has been revised and improved for the time normalization [Bethard and Parker, 2016] and to comply with international standards, which resulted in ISO-TimeML [Pustejovsky *et al*., 2010]. The ISO-TimeML standard is used for the annotation of different corpora and references in different languages. Examples of such corpora include: [Bittar, 2010] which is the building of a Time Bank for the French language, [Goel *et al*., 2020] a Time Bank for Hindi, the application of TimeML to Korean [You *et al*., 2011], etc.

TimeML aims to tackle the problem of recognizing an event and its temporal anchoring in a text. The temporal data in TimeML is annotated with the TIMEX3 tag which is inspired by TIMEX2 and uses most of its elements. However, in the TimeML project there is a binding between events and temporal data, an anchoring that is not considered in TIMEX2. Regarding their use cases, both TIMEX2 and TIMEX3 are intended for human annotators, but can also be used for the development of computer applications dedicated to the extraction and annotation of temporal data [Mani *et al*., 2001].

3. Introducing the TimeInfo annotation framework

While the TIMEX2 and TIMEX3 annotation systems provide a rich description of temporal data and aim for a general use of temporal data annotation. We propose a new temporal information annotation framework, called TimeInfo, which is specifically designed for the semantic categorization of temporal data in scientific papers. The main purpose of TimeInfo is to make possible building representations of temporal data in texts that convey as much as possible the linguistic meaning of a complex temporal expression.

In terms of semantics, TimeInfo includes most of the information that is provided by the previous systems (TIMEX2 and TIMEX3) and introduces some new attributes that allow for finer distinctions between temporal expressions and a richer representation. Figure 2 presents a diagram of TimeInfo, showing all different elements and all possible values that an attribute can take. In red, we have represented the attributes that are present in elements of TIMEX2 or TIMEX3. In blue, we

have represented the new attributes that are specific to our TimeInfo annotation framework.

Unlike TIMEX2 and TIMEX3, TimeInfo provides the possibility to represent complex temporal expressions, such as “from December 2001 to April 2002”, by recognizing them as intervals of time with their various attributes (granularity, duration, precision, etc.). Such an expression would be analysed by TIMEX3 as a text span having two dates which are “December 2001” and “April 2002”. In TimeInfo, the link between these two dates and the surrounding text is analysed to represent this temporal information as an interval having the following attributes:

interval = “closed”, granularity = “month”,
duration = 5, startDuration = “December 2001”,
endDuration = “April 2002”, indicator = “from-to”,
precision = “precise”, valType = “real value”.

To take another example, the expression “since the mid 1990s” would be represented by TimeInfo as:

interval = “ Right-open”, granularity = “year”,
indicator = “Since the”, precision = “imprecise”,
tempClue = “mid”, valType = “real value”

Thus, in addition to the information on dates and time, TimeInfo also relies on the linguistic context and syntactic elements that introduce the temporal data in the text to account for the different meanings that a temporal data can take. These syntactic elements are similar but not identical to the trigger elements that we find in [Ferro *et al*., 2003].

As can be seen in Figure 2, TimeInfo annotates temporal expressions with the TIMEINFO tag that can have 9 different attributes. An attribute can be mandatory, such as **granularity**, or optional, such as **duration**. Each attribute contains either a semantic value which allows us to identify the information carried by the temporal expression, or a syntactic indicator, which indicates linguistic and syntactic elements of the context and can take values from an open set. For example, we

can identify elements such as the beginning or the end of an event, represent actions that have ended, continue, or will end at a particular moment in time. In addition, we can identify if the temporal expression is precise, fuzzy, with temporal information presented as an asserted value or an estimate.

Hereafter, we describe the semantic role of each attribute. Then, at the end of this section we present examples of expressions annotated with TimeInfo (Figure 3).

3.1. Interval

The **interval** attribute is mandatory in the annotation schema. Indeed, interval carries the semantic value of the type of interval that is represented by the temporal data. It has three possible values: **closed**, **left-open** and **right-open**[1].

An interval is considered as closed if the expression allows us to identify both the **start** and the **end** of the temporal interval. Also, an interval with the **closed** value can describe a date, or a duration as shown in the following examples:

"***On January 2020**, SARS-CoV-2 was isolated and announced as a new, seventh, type of human coronavirus*."[Bzówka *et al*.,]

"*The surveys were conducted **from February 1, 2020 to February 10, 2020**, as transmission of COVID-19 peaked across China and stringent interventions were in place*."[Zhang *et al*., 2020]

To illustrate the **right-open** and **left-open** intervals, we can consider a representation of time as a straight line that goes from point **A** to point **B** (see Figure 1). In a **right-open** interval, the value of point **A** is known, and the value of point **B** is not known. For example:

1 From a set-theoretic point of view, four types of intervals exist: *closed, left-open, right-open* and *open*. For our annotation scheme we have considered only three of these types of intervals, leaving out the *open* intervals. In fact, we have not been able to observe any occurrences of open intervals in our datasets of scientific publications.

"*In Asia, several media outlets have opted to use ""Wuhan-pneumonia"" 7 instead of COVID-19 in their reporting even though WHO has explicitly advised against naming new human infectious diseases with geographic locations or populations* ***since 2015***."[Lin, 2020].

In a **left-open** interval, the starting point **A** is unknown, or not identifiable from the linguistic expression, and the end point **B** is known. For example:

"*Based on epidemiological data* ***before 2019****, only six CoVs proved to cause human respiratory diseases: i) HKU1, HCoV-NL63, HCoV-OC43 and HCoV-229E only lead to mild upper respiratory disease, but rarely bring about severe diseases in people; ii) SARS-CoV and MERS-CoV attack lower respiratory tract and always induce severe respiratory syndrome*."[Kang *et al*., 2020]

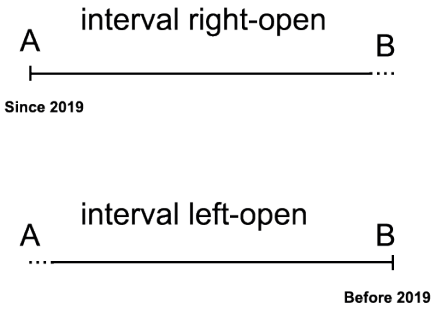

Fig. 1 – *Interval right-open and left-open*

3.2. Granularity

The second attribute that we introduce is **granularity**, and it is mandatory. The **granularity** represents the smallest unit of temporal division that is intended by the author of the expression. It takes as value one of the following: "**day**", "**month**", "**year**", "**decade**", "**century**", "**millennium**". For example, in the expression "**12 January 1992**", the value of **granularity** is **day**. By default, the smallest unit we use is "**day**", and this choice was motivated by our corpus which deals with

SARS-CoV and SARS-CoV-2. Smaller units can be considered for other datasets.

3.3. Duration, startDuration and endDuration

In addition to the **granularity**, we have the **duration** attribute which gives the number of days, months, years that span the temporal interval. For instance, in the expression **"from January 2021 to August 2021**" the value of **granularity** is **month,** and the **duration** is **8**. The information on the duration of an event can be useful in the context of information retrieval or for data visualization purposes. The **startDuration** and **endDuration** attributes provide the beginning and the end of a temporal interval. Examples are presented on figure 3.

3.4. Indicator

The **indicator** attribute stores the linguistic and syntactic elements (expressions) that introduce temporal data. These linguistic and syntactic elements are extracted from the context. The values of the **indicator** attribute are not limited to a closed set, but can be any linguistic expression, or a list of expressions that introduce the temporal data in the text, such as prepositions, e.g. "from … to", or adverbial phrases like "towards the end of". The presence of the INDICATOR attribute is intended to facilitate the process of the construction of algorithms for the annotation of temporal expressions. The syntactic elements that are given by the INDICATOR attribute can be used either as features for machine learning algorithms, or to develop linguistic resources and rules for the detection and annotation of time expressions.

3.5. Precision and tempClue

Some linguistic expressions indicate temporal data for which the boundaries (start and end of an event) cannot be precisely identified. For example, the expression "**in the mid-19th century**" points to a "fuzzy" interval when no specific year can be considered as a start or an end of the interval.

The value of precision can be **precise** as in the expression "*October 14, 2003*" or **imprecise** as in "*early December* ".

When the temporal expression is **imprecise**, it is often introduced by an adjective such as "early", "mid", "late", etc. The **tempClue** attribute stores such adjectives that point to parts of the fuzzy interval. For example, for the expression "in the mid-19th century", the value of **tempClue** is **mid**. Both **precision** and **tempClue** attributes may serve as a feature in a sophisticated search engine where a **precise** temporal data is sought. For example, the query: "Covid-19 cases before the end of July 2021" should, theoretically, retrieve data with mentions like "early 2021", or "in the early 2020s".

3.6. ValType

While analysing our corpus, we identified two different types of temporal expressions: estimated and real value. For example, in the expression "*the estimated possible date of emergence was January 12, 1992*" the temporal information is an estimated value, while the expression "*the date of emergence was January 12, 1992*" states the real temporal value. The **valType** attribute expresses this distinction. It is optional and can take two values: estimated value and real value. The purpose here is to separate the real values from the estimated values and this information is quite relevant from the perspective of information retrieval.

In figure 3, we present three examples of sentences from our corpus. These examples are annotated with TimeInfo.

4. Applications

As described above, this research aims to provide a unified framework for the annotation of temporal information in scientific papers taking into consideration the semantic dimension of the temporal expressions. The processing of temporal information is a cornerstone for many applications around the data mining of scientific papers, e.g., the creation of a search engine for scientific corpora that support temporal data requests. For instance, given a request as 'SARS-CoV virus

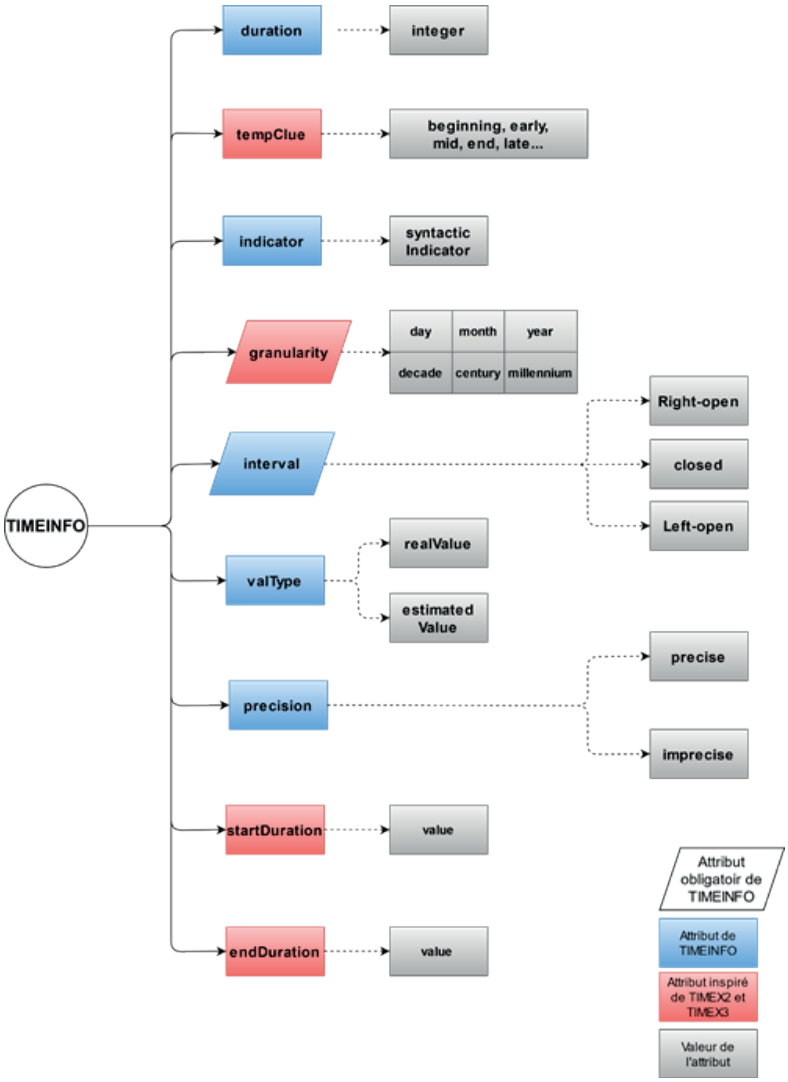

FIG. 2 – *Temporal Information Annotation Framework: TimeInfo*

before 01/01/2019.', the results would include the sentences and paragraphs extracted from scientific papers that report on research about the SARS-CoV virus before 01/01/2019.

By the aggregation of such data, timeline visualizations of research topics and research results can be produced automatically. Such tools would allow for the efficient large-scale analysis of corpora through the exploitation of their temporal data to help create states of the art, but also identifying topics and subjects that lack sufficient information and need more research.

Example 1 :

"<TimeInfo interval = "closed" granularity = "day" duration = "52" startDuration = "25 th December 2019" endDuration = "15 th Feb 2020" indicator = "from" valType = "realValue" precision = "precise">**From 25 th December 2019 to 15 th Feb 2020**</TimeInfo>, a total of 110 patients (45.5% female, mean age 64.03±16.54 year old) with suspected (n=30, 27.3%) or confirmed (n=80, 72.7%) COVID-19 were admitted in department of respiration or emergency department of Wuhan No.1 Hospital." [Zhang et al., 2020]

Example 2 :

"<TimeInfo interval = "closed" granularity = "month" duration = "1" tempClue = "mid" indicator = "in" valType = "realValue" precision = "imprecise">**In mid-February 2020**,</TimeInfo> the first clusters of 2019-nCoV emerged in northern Italy, near the southern border of Switzerland."
[Papachristofilou et al., 2020]

Example 3 :

"First, the empirical data that previous studies used were collected<TimeInfo interval = "left-open" granularity = "day" duration = "1" indicator = "before" valType = "realValue" precision = "precise">**before 25 th Jan, 2020.**</TimeInfo> "
[Li et al., 2020]

FIG. 3 – *Examples annotated with TimeInfo*

5. Conclusion

Our research aims to offer a unified framework and a methodology for the semantic representation and annotation of temporal data in full text scientific papers. Thereby, we have developed a new annotation framework for the semantic categorization of temporal expressions that outperforms existing annotation schemes by taking into consideration complex temporal expressions and allows for more fine-grained analysis.

The next step of our work will be the development of a tool (TimeInfo Tagger) that automatically extracts and annotates temporal data in scientific papers using our annotations scheme TimeInfo. Rather than relying on Named Entity Recognition for the extraction of temporal data, we intend to use sets of rules that can identify simple and complex temporal information and consider its linguistic context. Then, TimeInfo Tagger will be used for the generation of annotated scientific corpora.

References

Bethard and Parker, 2016. Bethard, S. and Parker, J. (2016). A semantically compositional

annotation scheme for time normalization. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3779-3786.

Bittar, 2010. Bittar, A. (2010). Building a TimeBank for French: a reference corpus annotated according to the ISO-TimeML standard. PhD thesis, Paris 7.

Bzówka *et al*.,. Bzówka, M., Mitusińska, K., Raczyńska, A., Samol, A., Tuszyński, J. A., and Góra, A. Structural and evolutionary analysis indicate that the sars-cov-2 mpro is an inconvenient target for small-molecule inhibitors design.

Chinchor, 1998. Chinchor, N. A. (1998). Overview of muc-7/met-2. Technical report, Science Applications International Corp San Diego CA.

Ferro *et al*., 2003. Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2003). Tides: 2003 standard for the annotation of temporal expressions. Technical report, MITRE Corp MClean Va Mclean.

Goel *et al*., 2020. Goel, P., Prabhu, S., Debnath, A., Modi, P., and Shrivastava, M.(2020). Hindi timebank: An iso-timeml annotated reference corpus. In 16th JointACL-ISO Workshop on Interoperable Semantic Annotation PROCEEDINGS, pages 13-21.

Kang *et al*., 2020. Kang, S., Peng, W., Zhu, Y., Lu, S., Zhou, M., Lin, W., Wu, W., Huang, S., Jiang, L., Luo, X., *et al.* (2020). Recent progress in understanding 2019 novel coronavirus (sars-cov-2) associated with human respiratory disease: detection, mechanisms and treatment. International journal of antimicrobial agents, 55(5):105950.

Lin, 2020. Lin, L. (2020). Solidarity with china as it holds the global front line during covid-19 outbreak.

Mani *et al*., 2001. Mani, I., Wilson, G., Ferro, L., and Sundheim, B. M. (2001). Guidelines for annotating temporal information. In Proceedings of the first international conference on Human language technology research.

Pustejovsky *et al*., 2010. Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). Iso-timeml: An international standard for semantic annotation. In LREC, volume 10, pages 394-397.

Saurí *et al*., 2006. Saurí, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (2006). Timeml annotation guidelines. Version, 1(1):31.

You *et al*., 2011. You, H.-J., Jang, H.-Y., Jo, Y.-M., Kim, Y.-S., Nam, S.-H., and Shin, H.-P. (2011). The korean timeml: A study of event and temporal information in korean text. Language and Information, 15(1):31-62.

Zhang *et al*., 2020. Zhang, J., Litvinova, M., Liang, Y., Wang, Y., Wang, W., Zhao, S., Wu, Q., Merler, S., Viboud, C., Vespignani, A., *et al.* (2020). Age profile of susceptibility, mixing, and social distancing shape the dynamics of the novel coronavirus disease 2019 outbreak in china. medrxiv.

Résumé

Dans cet article, nous proposons un nouveau schéma pour l'annotation des expressions temporelles dans les textes que nous nommons TimeInfo. Aussi, nous présentons et analysons des modèles existants pour l'annotation des données temporelles. Ensuite, nous définissons l'ensemble des éléments et leurs attributs dans TimeInfo et expliquons comment ces éléments fournissent des analyses plus fines pour l'annotation des expressions temporelles. L'objectif principal de notre modèle est de permettre une meilleure détection, extraction et représentation sémantique des données temporelles dans les articles scientifiques. Enfin, nous présentons les cas d'utilisation et les différentes applications qui peuvent être développées à l'aide du framework TimeInfo.

Mots-clés : TimeInfo · Information temporelle · Annotation sémantique · Fouille de textes scientifiques · Fouille de textes · Extraction d'information.