Annotating Temporal Information in Clinical Notes for Timeline Reconstruction: Towards the Definition of Calendar Expressions

Natalia Viani
IoPPN, King's College London
London, UK

Hegler Tissot
University College London
London, UK

Ariane Bernardino
IoPPN, King's College London
London, UK

Sumithra Velupillai
IoPPN, King's College London
London, UK
EECS, KTH, Stockholm, Sweden

Abstract

To automatically analyse complex trajectory information enclosed in clinical text (e.g. timing of symptoms, duration of treatment), it is important to understand the related temporal aspects, anchoring each event on an absolute point in time. In the clinical domain, few temporally annotated corpora are currently available. Moreover, underlying annotation schemas - which mainly rely on the TimeML standard - are not necessarily easily applicable for applications such as patient timeline reconstruction. In this work, we investigated how temporal information is documented in clinical text by annotating a corpus of medical reports with time expressions (TIMEXes), based on TimeML. The developed corpus is available to the NLP community. Starting from our annotations, we analysed the suitability of the TimeML TIMEX schema for capturing timeline information, identifying challenges and possible solutions. As a result, we propose a novel annotation schema that could be useful for timeline reconstruction: CALendar EXpression (CALEX).

1 Introduction and Background

When applying natural language processing (NLP) methods to the analysis of clinical notes, understanding the temporal aspects of narratives is crucial (e.g. *when* the patient experienced a certain symptom, or *when* a particular drug was prescribed). To model and extract the temporal information enclosed in free text, the development of suitable annotation schemas and reliably annotated corpora is essential.

The TimeML specification language was developed to enable the recognition of events and their temporal ordering in general-domain texts (Pustejovsky et al., 2003a). In the original schema, four major elements are modelled: time expressions (TIMEXes), events, signals, and their relations. Signals are function words (e.g. "during", "before") that indicate how temporal objects can be related to each other. Relations are represented by either temporal links (e.g. "before", "simultaneous"), subordination links (e.g. "intensional", "factive"), or aspectual links (e.g. "initiates, "continues"). The TimeML schema was used to develop the TimeBank corpus, consisting of 183 news articles (Pustejovsky et al., 2003b). Gold annotations were reused in the TempEval tasks on temporal information extraction (Verhagen et al., 2007; Pustejovsky and Verhagen, 2009), where a simplified TimeML annotation was applied.

The TimeML specification language provides a standard model for the mark-up of time expressions (with type Date, Time, Duration, or Set), events (mostly verbs or noun phrases), and their temporal ordering (Pustejovsky et al., 2010), and it can be in principle applied to any type of text. In the clinical domain, two reference corpora based on TimeML are available. The 2012 i2b2 corpus (310 discharge summaries) includes annotations for time expressions, clinical events, and eight types of temporal relations (Sun et al., 2013a). In addition, a section time (`SECTIME`) is used to keep track of section creation dates. The THYME corpus (1,254 oncology notes) contains annotations for events, time expressions (with two additional types), and 5 types of temporal relations (Styler IV et al., 2014). In this corpus, narrative containers were introduced, representing temporal buckets (mostly dates) containing a set of events. Figure 1 provides a graphical representation of the main changes introduced by i2b2 2012 and THYME on the original TimeML model.

Most clinical NLP development based on available corpora have focused on the three separate main tasks: detecting and classifying 1) events and

Proceedings of the BioNLP 2019 workshop, pages 201–210
Florence, Italy, August 1, 2019.

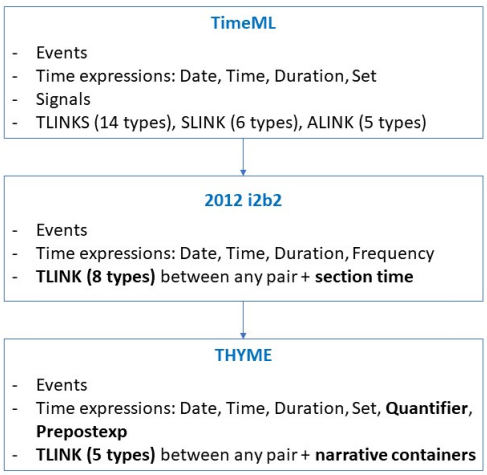

Figure 1: The TimeML model and how it has been adapted to, and implemented in, the clinical domain in the 2012 i2b2 and THYME corpora.

2) time expressions (and their attributes), and 3) classifying temporal links (TLINKs) (Sun et al., 2013b; Bethard et al., 2016, 2017). However, these separate tasks do not directly address the problem of anchoring events on an absolute timeline, which would be important for an improved understanding of patient trajectories. Moreover, mainly due to the inherent complexity of identifying temporal links (which can exist between any pair of entities and with different types), temporal annotation becomes a challenging task.

A few studies have proposed alternate approaches or extensions to the TimeML model for temporal ordering of events in various domains (Chambers et al., 2014; Jeblee and Hirst, 2018; Kolomiyets et al., 2012; Raghavan et al., 2014), but without addressing *timeline reconstruction*, i.e. anchoring events in absolute time. An approach that aimed to anchor events in time and simplify the annotation task was proposed by Reimers et al. (2016, 2018), where the event time is modelled as an argument of the event mention. However, the emphasis lies on the events, not on time expressions. Another approach proposed by Zhao et al. (2012) focuses on an alternative way to normalise time expressions using time intervals, allowing for more efficient temporal reasoning.

The existing temporally annotated corpora for the clinical domain have not, to our knowledge, been studied in great detail with respect to timeline reconstruction, particularly as regards TIMEX annotations. Tissot et al. (2015) found a surprising number of repeated inconsistencies between the guidelines and the manually created corpora for certain regular and unambiguous temporal language constructs. These were mainly related to inconsistencies in span and class assignments, incorrect annotations (false positives) or missing annotations (false negatives). This evidences how hard it is to create coherent annotated resources, and how NLP development and evaluation can be affected by the quality of the underlying data.

To further support the growth of clinical-temporal NLP and the development of translational applications, the release of additional annotated corpora (including different clinical specialties) is needed. Moreover, despite the efforts put into creating resources like the i2b2 2012 and THYME corpora, the suitability of the underlying annotation schemas to support clinical timeline reconstruction has not been widely investigated. While some types of TIMEXes are definitely useful to anchor clinical events on a timeline (e.g. explicit references to the calendar, like *in February 2009*), the importance of capturing other entities (e.g. the frequency of medication intake) remains unclear. Moreover, there might be TIMEX types that could be relevant in the clinical domain, but are not currently considered by temporal models based on TimeML, e.g. age-related expressions. As another important point, the way in which TIMEXes are typically normalised is not necessarily optimal for timeline reconstruction (e.g. duration values such as `"P4M"` cannot be immediately placed on a timeline).

Our intuition is that, by changing the way time expressions are defined and normalised, temporal relation annotation could be much simplified. More specifically, adding timeline information at the time expression level could help to temporally anchor entities without the need for multiple types of temporal links. To assess this, we applied a TimeML-based TIMEX annotation schema on a corpus of clinical texts, investigating how time expression information is documented, and then analysed how it can be reused for timeline reconstruction. As a result, we created a corpus of clinical texts (for four different clinical specialties), annotated with TIMEXes which mostly relies on the TimeML schema. This corpus is publicly available, as an additional resource that could be reused for temporal NLP.[1] In this paper, we analyse the suitability of these TIMEXes for timeline recon-

[1] https://github.com/medesto/timeline-reconstruction

struction, and from this analysis propose an innovative way to annotate temporal information in free text: CALendar EXpressions (CALEX).

2 Materials and Methods

2.1 Dataset

We downloaded and extracted documents from MTSamples,[2] a collection of Medical Transcription Sample reports for multiple clinical specialties (where the same document can belong to different groups), created for educational purposes and for working transcriptionists.

In the MTSamples resource, there is no availability of document creation times (DCTs). Most documents follow a semi-structured format, including different section headings/textual content depending on the specific clinical specialty. We selected the following specialties for manual annotation and analysis: discharge summaries (108 documents), psychiatry-psychology (53 documents), paediatrics (70 documents), and emergency (75 documents).[3]

2.2 Manual Annotation

All MTSamples subsets were double-annotated for five types of time expressions: Date, Time, Duration, Frequency (from TimeML (Pustejovsky et al., 2010)), and Age-related (*at the age of 16*, *in his teens* (Viani et al., 2018)). Identified expressions were also normalised to a standard temporal value (e.g. `"2011-05"` for *May 2011*). Manual annotations were performed by two annotators: a native English-speaker undergraduate student and a non-native English-speaker researcher. To guide the annotation process, we created specific annotation guidelines, which we refined by adding relevant examples from the text. Resulting guidelines are available on our repository.

Besides including an Age-related time expression type, our annotation task differed from TimeML in two ways. First, we included as TIMEXes domain-specific (and temporally-anchored) concepts, e.g. *On the day of admission*. This is similar to the Prepostexp type used in the THYME corpus, but instead of creating a different category, we included these expressions among existing types (e.g. Date). Depending on the clinical specialty (and more generally, on the domain), different concepts could be considered as temporal anchors within temporal annotation (e.g. *discharge* for discharge summaries, *pregnancy* for paediatrics notes). Second, we allowed annotators to use relative values in the normalisation phase, if needed. These relative values, or formulas, can either refer to the document creation date (e.g. `"DCT-P2D"` for *Two days ago*) or to the domain-specific concepts (e.g. `"OP+P2D"` for *postoperative day #2*).

To compute inter-annotator agreement (IAA) on textual spans, we used the F1 score (allowing overlapping annotations). Expressions identified by both annotators (*overlap*) were used to compute IAA on types and normalised values (accuracy). For IAA on types, we also report the Cohen's Kappa measure (`K`), to take the possibility of chance agreement into account.

2.3 Annotation Analysis for Timeline Reconstruction

To analyse the suitability of using the TimeML-based TIMEX annotations for timeline reconstruction, we based our analysis on the following: 1) timeline properties of the TIMEX type Frequency, 2) properties of normalised values for Date annotations, and 3) properties of common annotation disagreements.

Our hypothesis regarding Frequency annotations was that these would not be necessarily useful as temporal references on an absolute timeline, as they would be mostly related to drug prescriptions. To assess this, we applied the MedEx-UIMA tool (Jiang et al., 2014) on the text surrounding each Frequency expression (the *context*),[4] and quantified the proportion of annotations close to drug mentions.

Normalised values for Date can represent a specific point on a timeline, e.g. `"YYYY-MM-DD"`, `"YYYY-MM"` or `"YYYY"`. However, they can also represent other less straightforward points in time, e.g. DCT-related (*yesterday*), vague references (*in the past*), incomplete dates (*on the 13th*), and concept-related (*on the day of admission*). To better understand these latter types of normalised Date values and their relation to timeline reconstruction, we analysed annotations marked as such by at least one annotator.

Finally, to inform the development of a new annotation schema for timeline reconstruction based

[2] www.mtsamples.com

[3] The number of unique documents is 286.

[4] We considered a window of 50 characters before and after the annotation.

on calendar expressions, we counted all annotation disagreements and analysed the most common type. We manually reviewed the documents containing these expressions, assessing whether: a) they could be placed on a timeline; and b) how they could be normalised in a non-ambiguous way. During this review, we also added new types of expressions that we believe would be crucial elements to anchor on a timeline, thus forming a proposal for a novel annotation schema.

3 Results

We report results for the new TimeML-based TIMEX annotated corpus in terms of IAA, and a breakdown of the number of documents, tokens and TIMEXes for each clinical specialty (Table 1). Furthermore, we report the results for the analysis on different aspects of the suitability of these annotations for timeline reconstruction that was used to inform the development of a novel annotation schema (further outlined in Section 4).

3.1 Manual Annotation

For each clinical specialty, Table 1 reports the number of documents (with total number of tokens), the number of time expressions marked by at least one annotator (*merged*), and those marked by both annotators (*overlap*). For IAA, we report F1 score for text spans (allowing overlapping annotations) and type/value agreement measures (on *overlap* annotations). We also report the prevalence of time expression types (only looking at overlap annotations with type agreement).

IAA results for text spans are encouraging, 76-84%. We observe that the distribution of TIMEX types is similar across clinical specialties, where Date is most common (28-36%) and Time is least common (3-9%), with the exception of discharge summaries, where Frequency is most common (39%) – probably due to the abundance of drugs prescribed after discharge. Agreement for normalised values measured by accuracy is slightly lower, overall (72-75%).

3.2 Annotation Analysis for Timeline Reconstruction

As shown in Table 1, Frequency expressions are common across all MTSamples subsets. By applying MedEx-UIMA and extracting the related contexts, we found that most frequencies occurred close to a drug mention (94%, 82%, 59%, and 80%, in discharge summaries, psychiatry-psychology, paediatrics, and emergency, respectively). By manually reviewing a sample of the remaining expressions, we noticed that some of them referred to alcohol/smoking (*he drank one bottle of wine everyday*) or recommendations (*continue bathing twice a week*), and would therefore not be placed on a timeline. In other cases, examples were still related to a drug mention which, however, did not fall in the selected context or was not extracted by MedEx.

As regards the analysis of Date normalised values, we noticed that most "non-standard" values were given by DCT-related formulas (e.g. `"DCT-P2Y"`). For discharge summaries, the second most frequent type was concept-related (e.g. `"ADM+P2Y"`, where ADM stands for ADMission day). Vague values were used across all subsets to mark time references that were not explicitly written (*at that time*, *on the following Tuesday*).

In all MTSamples subsets, the most frequent type of disagreement was Duration-vs-Date, with a proportion of 47% over all other types of disagreements (41/86, 51/98, 21/59, 34/69, in discharge summaries, psychiatry-psychology, paediatrics, and emergency, respectively). In Table 2, we report the most common types of disagreement across all subsets.

By manually analysing documents where these disagreements were present, and taking into considerations our findings on the (TimeML-based TIMEX) annotated corpus, we propose a new annotation schema for capturing time expressions that are actually useful for timeline reconstruction: CALendar Expressions (CALEX).

4 CALEX

CALEX refers to a temporal annotation schema restricted to time expressions and concepts that can be (directly or not) connected to an absolute timeline. The key novelty of this model is to better utilise time expression properties that are relevant for anchoring points on a timeline, including the introduction of certain timeline-relevant concepts.

In relation to TimeML-based TIMEX definitions, CALEX *excludes* the following, because they cannot be directly used for timeline reconstruction:

- FREQUENCY/SET/QUANTIFIER, e.g. *once a week*, *two units of blood*;

	dis. summaries	psych.	paediatrics	emergency
Documents	108	53	70	75
Tokens	55,433 (513/doc)	67,569 (1275/doc)	36,675 (524/doc)	52,041 (694/doc)
TIMEXes (merged)	1,378	1,227	566	801
TIMEXes (overlap)	994	840	360	496
TIMEXes (same type)	908	742	301	427
Date	326 (36%)	234 (32%)	85 (28%)	154 (36%)
Duration	110 (12%)	122 (16%)	49 (16%)	44 (10%)
Time	29 (3%)	31 (4%)	23 (8%)	39 (9%)
Frequency	355 (39%)	216 (29%)	61 (20%)	88 (21%)
Age_related	88 (10%)	139 (19%)	83 (28%)	102 (24%)
IAA F1	0.84	0.81	0.78	0.76
type acc.	0.91	0.88	0.84	0.86
type K	0.89	0.85	0.79	0.82
value acc.	0.74	0.72	0.75	0.74

Table 1: Manual annotation results - time expressions (TIMEXes) on documents from MTSamples: discharge summaries (dis. summaries), psychiatry/psychology (psych.), paediatrics and emergency department documents.

Type	dis. summ.	psych.	paediatrics	emergency
Duration-vs-Date	41	51	21	34
Duration-vs-Time	11	5	8	16
Duration-vs-Frequency	16	5	7	6
Age_related-vs-Date	1	10	11	1
Age_related-vs-Duration	1	7	8	2
Date-vs-Time	2	6	2	6
Frequency-vs-Time	6	7	1	1

Table 2: TIMEX type disagreement counts on the MTSamples subsets - discharge summaries (dis. summ.), psychiatry/psychology (psych.), paediatrics and emergency department documents.

- DURATION when it is a temporal *attribute* describing other events, e.g. "the procedure usually takes *15 minutes*";
- TIME when it refers to temporal *attributes* describing other events, e.g. "to be always taken *around 9am*".

There are three main elements in the proposed CALEX annotation schema: `TYPE`, `METADATA`, and `VALUE`.

TYPE

`TYPE` defines the type of a calendar expression. The possible types within the CALEX schema are described as follows:

- `CALENDAR`: this type covers all calendar expressions that do not require any metadata in order to provide the final normalised `VALUE`, including:
  - explicit calendar references in different temporal granularities such as date, month, year
  - timestamps
  - explicit ranges
  - when time is described as a *period* of time (duration) but the connection with the timeline is not clear or explicit - this type refers to the original DURATION type as part of the TimeML annotation guidelines (e.g. "he took this medication *for one month*")
- `AGE`: age-related expressions can either define the current age of a patient (e.g. "a *56 year old* woman"), or be a reference to a certain point in time in which the patient had a given age (*at the age of 17*).
- `DOMAIN`: expressions that either explicitly define the value of a domain-specific concept

(*admitted on 2010 Jun 6th*), or are references to a given domain-specific concept (*on post-operative day #4*).

- `DCT`: expressions that require information about the document creation time in order to be normalised (*last month*).
- `TENSE`: imprecise expressions that refer to conditions in the past, present or future (*recently*).
- `CONTEXT`: expressions that refer to a *temporal context*, represented by either the last mentioned temporal reference or the most recent temporal reference available within the document. [5] This type includes times/periods of the day where the connection to the timeline is not clear and relies on the temporal context (e.g. *the previous night*). However, times/periods of the day representing frequencies (e.g. "one tablet *at night*") are NOT considered as CALEXes.

METADATA

We introduce `METADATA` as a feature to allow for a computationally more efficient way of calculating a particular time reference for CALEXes that are not explicitly anchored in time. An essential aspect of this feature is that it can include concepts in its definition. These can also be explicitly set within the `METADATA` feature, to ensure the original values are used in order to normalise the final calendar expression.

Document-related concepts include the document creation time {`doc.DCT`} and contextualised references to the last or more recent temporal mentions within the text ({`doc.LAST`} and {`doc.RECENT`}). Patient-related concepts are used to describe patient demographic features, such as {`patient.AGE`}, {`patient.DOB`}, {`patient.DOD`}, the later possibly useful when analysing death certificates.[6] Patient-stay-related concepts will basically refer to the period within admission and discharge ({`patient.ADMISSION`} and {`patient.DISCHARGE`}).

One important concept that may require some disambiguation is related to the pregnancy period. The terms *Pregnancy* and *Prenatal* are generally interchangeably used when referring either to the mother or the child. We formalise *Pregnancy* as being the period of time used when referring to the mother as a patient, whereas *Prenatal* refers to the period of time (usually 40 weeks) before {`patient.DOB`}, which refers to the child as a patient. This way, {`patient.PREGNANCY`} can occur at any time in the patient's life, whereas {`patient.PRENATAL`} is the period of 40 weeks preceding the patient's date of birth.

Finally, some social- and family-related concepts can be used in order to refer to some temporal values regarding the patient's relatives, such as {`mother.AGE`} or {`father.DOB`}.

Besides making use of timeline-relevant concepts, `METADATA` also contains functions that are used to derive values:

- `.set()`: for explicitly defining the value of domain-specific concepts;
- `.add()`: adds a period of time to a given point in the calendar, moving to a later point in time;
- `.sub()`: subtracts a period of time from a given point in the calendar, moving to an earlier point in time;
- `.next()`: finds the next occurrence of a temporal granularity based on an anchor calendar expression;
- `.prev()`: finds the previous occurrence of a temporal granularity based on an anchor calendar expression.

For example, a reference to DCT cannot be properly normalised when DCT is unknown. However, the metadata can keep the definition for a calendar expression, to be converted to an actual value when DCT is given: instead of parsing the entire document, only the metadata has to be re-evaluated – e.g. metadata for the expression "yesterday" is "{`doc.DCT`}`.sub(P1D)`".

VALUE

This component gives a normalised value of a calendar expression, mostly following the previous TimeML notation, with an extension: *range* values are used to normalise periods of time in the form of [begin,end].[7]

[5] Other specific contextual references can be required for documents in different domains.

[6] {patient.AGE} and {patient.DOB} represent complementary concepts.

[7] To indicate included endpoints, we use standard square brackets: [A,B]. To indicate excluded endpoints, we use re-

Example	Type	Metadata	Value
dated June 15, 2007	CALENDAR	null	2007-06-15
on June 15, 2007 at 10:00	CALENDAR	null	2007-06-15T10:00
in 2009	CALENDAR	null	2009
between 2007 and 2009	CALENDAR	null	[2007,2009]
since 2007	CALENDAR	null	[2007,]
after 2007	CALENDAR	null	]2007,]
for one month	CALENDAR	null	[,P1M]
a 20-year-old male patient ...	AGE	{patient.AGE}.set(P20Y)	P20Y
at age 15, when...	AGE	{patient.DOB}.add(P15Y)	XXXX (unknown DOB)
at age 15, when...	AGE	{patient.DOB}.add(P15Y)	[2002-04,2003-03] (known DOB)
since age 25 ...	AGE	[{patient.DOB}.add(P25Y),]	[XXXX,] (unknown DOB)
admitted on 05-27-2009	DOMAIN	{patient.ADMISSION}.set(2009-05-27)	2009-05-27
born in 07/2007	DOMAIN	{patient.DOB}.set(2007-07)	2007-07
discharged on 01/21/10	DOMAIN	{patient.DISCHARGE}.set(2010-01-21)	2010-01-21
upon discharge	DOMAIN	{patient.DISCHARGE}	XXXX-XX-XX (unknown)
18 hours prior to admission	DOMAIN	{patient.ADMISSION}.sub(PT18H)	2010-06-25T02:00
tomorrow	DCT	{doc.DCT}.add(P1D)	2010-07-02
11 years ago	DCT	{doc.DCT}.sub(P11Y)	1999
for the next 2 weeks	DCT	[{doc.DCT},P2W]	[2010-07-01,2010-07-15]
next Tuesday	DCT	{doc.DCT}.next(WD,3)	2010-07-06
in july of next year	DCT	{doc.DCT}.add(P1Y).next(M,7)	2011-07
in the past	TENSE	[,{doc.DCT}[	[,2010-07-01[
recently	TENSE	],{doc.DCT}[	],2010-07-01[
at this time	TENSE	]{doc.DCT}[	]2010-07-01[
in the future	TENSE	]{doc.DCT},]	]2010-07-01,]
at that time	CONTEXT	{doc.LAST}	2010-03-15
3 days prior	CONTEXT	{doc.LAST}.sub(P3D)	2010-03-12
10am	CONTEXT	{doc.LAST}.next(TH,10)	2010-03-15T10:00
was...on Tuesday	CONTEXT	{doc.LAST}.prev(WD,3)	2010-03-09

* `doc.DCT = "2010-07-01"` for all the examples

Table 3: Calendar Expression — CALEX — examples.

Table 3 presents some examples on how the CALEX annotation schema works in terms of normalising the main features.[8]

As shown in the examples, a key element of the CALEX schema is the handling of domain-specific concepts in the `METADATA` element.

In Table 4, we show how different expressions would be represented within CALEX and TimeML, highlighting the types to be added to capture timeline-related expressions in CALEX format ("N/A" values in the *TimeML type* column).

Figure 2 provides an example of timeline creation using CALEX instead of TimeML (for the psychiatry domain). First, to temporally anchor the first emergence of auditory hallucinations, an age-related time expression is added (*since the age of 14*). Second, to capture the admission date, a specific domain concept is used (*On admission*, abbreviated as {patient.ADM}). For these expressions, the `METADATA` feature allows identifying a specific point in the timeline without the need for temporal links. As another difference, the medication frequency (*twice a day*), which cannot be represented at the timeline level, is removed.

verse square brackets:]A,B[. Open ranges/periods of time are indicated by [A,] or [,B].

[8]Note that relevant prepositions are included in the expression textual span.

5 Discussion

In this paper, we investigated how temporal information is documented in clinical text by focusing on time expressions (TIMEXes), using clinical notes from MTSamples for four different specialties (discharge summaries, psychiatry and psychology, paediatrics, and emergency). Our goal was to assess whether TIMEX annotation schemas based on TimeML would be suitable to capture the information needed to reconstruct patient timelines. First, we annotated documents using TimeML-inspired TIMEX types. Then, we analysed which of these expressions actually indicate a connection to the timeline, thus proposing a new annotation schema based on calendar expressions: CALEX.

Annotating MTSamples documents with a TimeML-based TIMEX model was helpful to investigate how temporal information is reported across different clinical specialties. Despite the use of sample reports, which might be more structured as compared to real clinical records, the distribution of time expression types (Table 1) is similar to those found in i2b2 2012 and THYME, where Date represents the most common TIMEX type and Time the least common. By analysing our manually annotated time expressions, we identified some key points to be taken into account to simplify timeline reconstruction. First, we ob-

CALEX type	Example	Definition	TimeML type
CALENDAR	{on 02/12/2009}	directly connected to calendar	Date
DCT	{tomorrow}	relative to the DCT	Date
TENSE	{in the past}	imprecise reference	Date
DCT	{two years ago}	relative to the DCT	Duration
DOMAIN	{18 hours prior to admission}	related to a domain concept	Duration
CONTEXT	{three days before}	related to another expression	Duration
—	the procedure usually takes {15 minutes}	not directly connected to calendar	Duration
CALENDAR	on 02/12/2009 {at 9am}	directly connected to calendar	Time
—	{twice a day}	any re-occurring expression	Frequency/Set/Quantifier
AGE	a {56 years old} woman	age of the patient	N/A
AGE	{when she was 17}	reference to age	N/A
DOMAIN	{admitted on Oct 12th}	domain-concept definition	N/A
DOMAIN	{the day before admission}	reference to domain	N/A

Table 4: Time expression examples as represented within CALEX and TimeML.

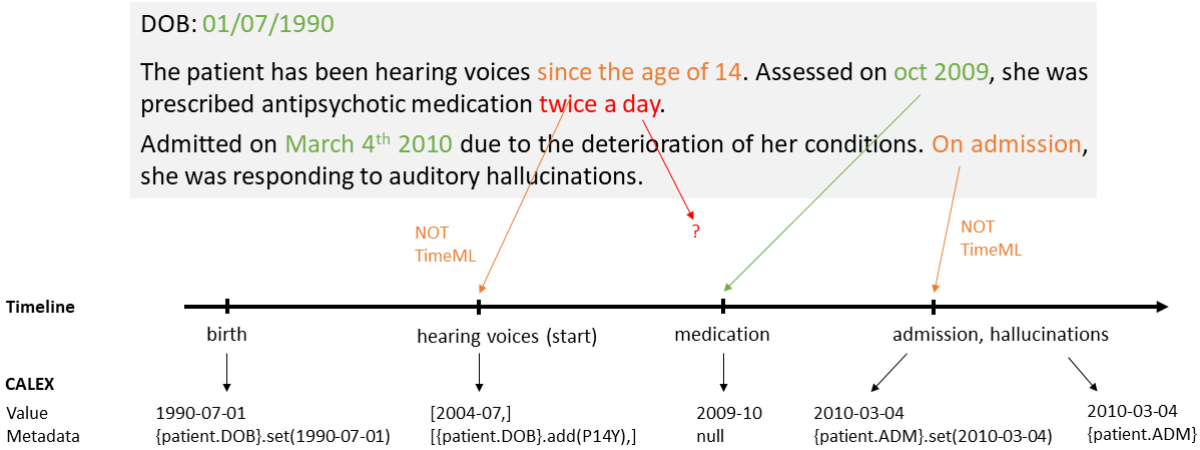

Figure 2: Example of timeline creation using CALEX instead of TimeML: an age-related and a domain-specific time expressions are added, while the medication frequency is removed.

served that most Frequency annotations are not helpful to anchor clinical events on an absolute timeline. Our suggestion is to remove such elements from time expression annotations, and to capture them as entity attributes instead (e.g. drug prescriptions). Moreover, the TimeML TIMEX normalisation step is not always directly useful for timeline reconstruction, as some expressions would still require different types of temporal links to be connected to the calendar.

To address these points, our proposed model, CALEX, integrates timeline information at the time expression level, specifying three different components: `TYPE`, `METADATA`, and `VALUE`. The new `TYPE` classification allows distinguishing between expressions directly connected to the calendar (e.g. full explicit dates) and relative/contextual expressions. Besides facilitating timeline reconstruction, this should also reduce ambiguity when assigning expression types, as the different type categories are more clearly separated. The `METADATA` feature, in combination with `TYPE`, allows storing the information needed for calendar normalisation, making use of functions and timeline-relevant concepts (Table 3). While functions are general and reusable across different document types, timeline-relevant concepts are specific to each domain or use-case, capturing the most appropriate anchor points within a finite set (e.g. {Admission, Discharge} for discharge summaries). The `METADATA` feature is useful to automatically derive or evaluate the normalised `VALUE`, especially for concept-related/contextual expressions where manually assigning values might be not straightforward.

Compared to TimeML, the CALEX model removes the Frequency/Quantifier/Set type and introduces new types and normalised values. In particular, Date- and Duration- like expressions are assigned different CALEX types depending on how they can be linked to the timeline (Table 4), which will be useful to reduce manual annotation for temporal links. Within the CALEX model, instead, a greater annotation effort is required for the normalisation task. Especially for relative expressions, assigning standardized values to `METADATA` is likely to be hard for non-technical annotators. Therefore decisions on what

to manually annotate in the CALEX model will need to be defined. As a first step, we would require manual annotations mainly for calendar expression VALUEs, specifying the METADATA feature only if necessary (e.g. when no is DCT available) and using a simplified notation (e.g. "DCT-P2Y"). In most cases, this feature would be derived programmatically, and its derivative value used for evaluating the manual VALUE.

This study has some limitations and directions for future work. The TimeML-based TIMEX annotations have not been adjudicated. However, we have released the corpus as it is, so that NLP researchers can integrate/reuse annotations for analysis and system development. In particular, we have made available all annotations (merged), specifying which ones are overlapping (and could therefore be considered as more reliable). Our study has been heavily focused on analysing time expressions: we have not systematically also analysed how existing annotation schemas can capture calendar information by other annotation elements. For example, i2b2 2012 and THYME include annotations for admission and discharge, but they are classified as *events* to be linked to other temporal entities. In other studies, it has been proposed to add timeline information directly as event attributes, e.g. Reimers et al. (2018).

When normalising time expressions, another aspect to be considered is the presence of imprecise temporal references, which are abundant in the clinical domain (Tissot et al. (2019)). As part of the CALEX model, TENSE expressions are included, which are used to refer to the past, present or future. At the moment, we are also evaluating how to incorporate other types of imprecise temporal references. More generally, we are designing a CALEX annotation guideline which is focused on both manual (e.g. VALUE) and potentially automatic (e.g. METADATA) tasks. As future work, we plan to create a reference corpus annotated with CALEXes, and design a shared task for further evaluation. Creating a CALEX annotated corpus will be crucial to assess the utility of our model, as well as to highlight potential issues and areas for improvement (with a specific focus on the proposed types and the METADATA feature).

6 Conclusion

In this paper we developed a corpus of medical reports annotated with TimeML-based time expressions and systematically analysed their usefulness for timeline reconstruction. As a result, we proposed a new annotation schema, CALEX, which will be used to design and develop new resources.

Acknowledgments

We thank the reviewers for valuable comments on our manuscript. SV has received support from the Swedish Research Council (2015-00359), and the Marie Skłodowska Curie Actions, Cofund, Project INCA 600398. HT is funded and supported by the Health Data Research UK (grant No. LOND1).

References

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2(1):273–284.

Serena Jeblee and Graeme Hirst. 2018. Listwise temporal ordering of events in clinical notes. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 177–182, Brussels, Belgium. Association for Computational Linguistics.

Min Jiang, Yonghui Wu, Anushi Shah, Priyanka Priyanka, Joshua C Denny, and Hua Xu. 2014. Extracting and standardizing medication information in clinical text–the medex-uima system. *AMIA Summits on Translational Science Proceedings*, 2014:37.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 88–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pustejovsky, José M Castano, Robert Ingria, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust

specification of event and temporal expressions in text.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003b. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

James Pustejovsky and Marc Verhagen. 2009. Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 112–116.

Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 998–1008, Baltimore, Maryland. Association for Computational Linguistics.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204, Berlin, Germany. Association for Computational Linguistics.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2018. Event time extraction with a decision tree of neural classifiers. *Transactions of the Association for Computational Linguistics*, 6:77–89.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46:S5–S12.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Hegler Tissot, Marcos Didonet Del Fabro, Leon Derczynski, and Angus Roberts. 2019. Normalisation of imprecise temporal expressions extracted from text. *Knowledge and Information Systems*.

Hegler Tissot, Angus Roberts, Leon Derczynski, Genevieve Gorrell, and Marcos Didonet Del Fabro. 2015. Analysis of temporal expressions annotated in clinical notes. In *Proceedings of 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 93–102, London, UK. ACL.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 75–80.

Natalia Viani, Lucia Yin, Joyce Kam, Ayunni Alawi, André Bittar, Rina Dutta, Rashmi Patel, Robert Stewart, and Sumithra Velupillai. 2018. Time expressions in mental health records for symptom onset extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 183–192, Brussels, Belgium. Association for Computational Linguistics.

Ran Zhao, Quang Do, and Dan Roth. 2012. A robust shallow temporal reasoning system. In *Proceedings of the Demonstration Session at the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 29–32, Montréal, Canada. Association for Computational Linguistics.