Time Expressions in Mental Health Records for Symptom Onset Extraction

Natalia Viani, Lucia Yin, Joyce Kam, André Bittar
IoPPN, King's College London
London, UK

Ayunni Alawi
University of Sheffield
Sheffield, UK

Rina Dutta, Rashmi Patel, Robert Stewart
IoPPN, King's College London
SLaM NHS Foundation Trust
London, UK

Sumithra Velupillai
IoPPN, King's College London
London, UK;
KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

For psychiatric disorders such as schizophrenia, longer durations of untreated psychosis are associated with worse intervention outcomes. Data included in electronic health records (EHRs) can be useful for retrospective clinical studies, but much of this is stored as unstructured text which cannot be directly used in computation. Natural Language Processing (NLP) methods can be used to extract this data, in order to identify symptoms and treatments from mental health records, and temporally anchor the first emergence of these. We are developing an EHR corpus annotated with time expressions, clinical entities and their relations, to be used for NLP development. In this study, we focus on the first step, identifying time expressions in EHRs for patients with schizophrenia. We developed a gold standard corpus, compared this corpus to other related corpora in terms of content and time expression prevalence, and adapted two NLP systems for extracting time expressions. To the best of our knowledge, this is the first resource annotated for temporal entities in the mental health domain.

1 Introduction and Background

For psychiatric disorders such as schizophrenia, prolonged periods of time without treatment are associated with worse intervention outcomes (Kisely et al., 2006). The number of days between first symptom onset and initiation of adequate treatment is defined as duration of untreated psychosis (DUP). For patients with schizophrenia, a longer DUP has been linked to poorer cognitive function at the time of first presentation (Lappin et al., 2007). In addition, it has been shown to predict more severe symptoms and greater social and functional impairment (Hill et al., 2012). Therefore, identifying and reducing the DUP could significantly improve both clinical and functional outcomes. Starting from this observation, there is an increasing interest in measuring the DUP across large clinical samples, to provide a quality measure for mental health services, and in developing international guidelines aimed at reducing this value, thus improving outcomes at different levels (Connor et al., 2013).

Routinely collected data from health services, such as electronic health records (EHRs), can be useful for large-scale retrospective clinical studies. In mental health services, a large proportion of clinically relevant information is recorded only in open text fields. To make this information available for computational analysis, Natural Language Processing (NLP) methods can be used (Meystre et al., 2008; Wang et al., 2018). When applying NLP techniques to the clinical domain, one crucial task involves the identification of *temporal information*. In general, for temporal information modeling, three different steps are typically outlined: (i) the identification of relevant concepts, such as symptoms (*hallucinations*) and treatments (*Clozapine*), (ii) the identification of time expressions (*May 1st*), and (iii) the identification of temporal relations between entity pairs ({hallucinations} BEFORE {Clozapine}).

Over the past years, methods for temporal information extraction have been developed with promising results, mainly based on the ISO-TimeML specification language that was devel-

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 183–192
Brussels, Belgium, October 31, 2018.

oped for the general NLP domain (Pustejovsky et al., 2010). In the clinical domain, a few manually annotated corpora (*gold standards*) have been created. As part of the Informatics for Integrating Biology and the Bedside (i2b2) project, a set of 310 de-identified discharge summaries from an Intensive Care Unit (ICU) were annotated with events, time expressions, and temporal relations (Sun et al., 2013a). This corpus was then used for organizing the 2012 i2b2 Challenge on temporal relation extraction (Sun et al., 2013b). In the oncology field, Styler IV et al. developed a corpus of 1,254 de-identified EHR notes, annotated for both clinical and temporal information (THYME corpus) (Styler IV et al., 2014). This corpus has been used in different NLP challenges, among which Clinical TempEval 2015 and 2016 focused on temporal information extraction (440 and 591 documents, respectively) (Bethard et al., 2015, 2016). In both these corpora, four main TimeML types of time expressions (TIMEXes) are defined: Date, Duration, Frequency (or Set), and Time. The THYME corpus also includes two additional TIMEX types specific to the oncology domain: PrePostExp (expressions indicating Pre- and Post-operational concepts) and Quantifier (expressions like *twice* or *four times*).

Compared to other clinical domains, mental health records are characterized by a greater extent of narrative portions, describing symptomology and health progression without relying on structured fields. In this framework, relevant temporal information (e.g., associated to symptom onset or treatment initiation) is not always well represented by current temporal models. For example, identifying expressions like *at age 8* or *in 3rd year of secondary school* is not straightforward, especially as regards the normalization phase (e.g., converting *6th May 2018* to "2018-05-06").

Our long-term goal is to accurately identify symptoms and treatments from mental health records, and anchoring these on a timeline, to be able to calculate DUP and other clinically relevant temporal constructs on a large patient cohort. To address this long-term goal, we are developing a corpus with annotations that cover all necessary elements (time expressions, clinical entities and their relations).

In this study, we focus on one subgoal: addressing the problem of accurately identifying time expressions in mental health records related to patients who have been diagnosed with schizophrenia. Our aim was (i) to develop a gold standard corpus with time expression annotations, (ii) to analyze and compare typical time expressions in this corpus with other clinical corpora that have been annotated with time information (i2b2 2012, Clinical TempEval 2016), and (iii) to perform a feasibility study on adapting existing NLP systems for extracting time expressions.

2 Materials and Methods

2.1 Data

In this study, we used anonymized[1] mental health records from the Clinical Record Interactive Search (CRIS) database (Perera et al., 2016)[2]. This database was derived in 2008 from the EHR system adopted by a large mental healthcare provider in southeast London: the South London and Maudsley National Health Service (NHS) Foundation Trust (SLaM).

Mental health records for patients who had received a diagnosis related to schizophrenia were extracted. To identify these patients, we queried the CRIS database for patients who had been documented with an ICD-10 code for this disease (F20*) or, if not documented with a structured code, we relied on the output of an NLP tool which extracts diagnoses from free text (based on the keyword "schizophrenia") (Perera et al., 2016), resulting in 8,483 documents for 1,691 patients[3]. To make the task feasible for manual annotation and relevant to the clinical use-case, two main document sample steps were taken:

1. Only documents that were written within 3 months of first presentation to mental health services were extracted, on the assumption that these early documents would most likely contain the richest description of the patient's clinical history and presenting complaints related to relevant symptoms;

2. From these documents, only the longest document (in terms of total number of characters) for each patient was extracted to be used for annotation, on the assumption that this

[1]Textual portions are automatically anonymized by removing patient and relative identifiers, such as names and postal codes.

[2]This database has ethical approval for secondary analysis (Oxford REC C, reference 08 H0606 71+5).

[3]Data were extracted on March 31st 2016 for patients accepted in services after January 1st 2012.

document most likely would contain most information about the patient history;

From the extracted set, a random sample of 52 documents (one document per patient) was used in the time expression annotation task for creating our corpus.

2.2 Time Expression Annotations

As a first step for the extraction of psychosis symptom onset, it is necessary to identify all the time expressions occurring inside the text (e.g., *May 2012*, *a year ago*). These expressions can then be used at a later stage, to link each mentioned symptom to the corresponding date or time.

To enable the development of an accurate temporal extraction system, we manually annotated the available corpus with occurrences of time expressions, marking both TIMEX spans and types (e.g., Date). To facilitate this task, we prepared domain-specific annotation guidelines, inspired by the guidelines used in the 2012 i2b2 challenge (Sun et al., 2013a) and the THYME project (Styler IV et al., 2014).

In addition, we performed a comparative analysis with existing corpora (i2b2 2012 and Clinical TempEval 2016), to highlight similarities and differences, and to gain deeper knowledge in domain-specific characteristics related to how time information is documented in clinical text.

Comparison to Related Corpora and Guidelines Adaptation

Both the i2b2 2012 and the Clinical TempEval 2016 corpora are characterized by relatively short notes, with content organized in semi-structured sections (e.g., "History of present illness", "Hospital course"). To develop guidelines tailored to the mental health domain, we manually reviewed a few example documents to identify initial domain-specific requirements. In our corpus, most documents have few or no systematic section, with clinical and temporal information scattered across many different paragraphs. Moreover, symptoms and their onset are frequently associated to vague dates, as opposed to most events documented within the ICU and the oncology domain (e.g., problems, exams, operations). As a consequence, we found that the examples included in the i2b2 2012 and THYME guidelines did not capture all the time expressions that are typical of the mental health domain, and we decided to adapt them in order to simplify and clarify the annotation task. First, we only kept the TIMEX types that were relevant to the considered clinical use-case[4]: Date (e.g., *in May 2012*, *yesterday*), Time (e.g., *in the morning*, *3 pm*), Duration (e.g., *for three years*, *over the past two weeks*), and Frequency (e.g., *daily*, *twice a week*). Within Dates, we explicitly included generic expressions such as *past* and *current*, to enable temporal contextualization of events that cannot be anchored to specific TIMEXes. As for Frequencies, we put a particular focus on medication-related TIMEXes and domain-specific expressions (e.g., *OD* for "once daily"). We also defined an additional TIMEX type for "age-related" expressions, to capture clinically relevant temporal patient information. Although this type is not included in common TimeML models, it has been previously investigated as it can encompass relevant temporal information in a clinical setting (Wang et al., 2016). In this study, besides looking at the patient's current age (e.g., *28-year-old* man), we included all the expressions that rely on the date of birth in order to be correctly normalized (see Section 3.1). The final guidelines, which were written and revised by two NLP researchers, describe: the annotation task and goal, the TimeML TIMEX types (with sentences taken from the reference guidelines), and the domain-specific TIMEX types and examples.

Annotation Process

Annotations were carried out by three medical students, using the eHOST annotation tool (South et al., 2012). The students were all native English speakers and in their 1st-3rd year of medical studies. The corpus was randomly divided into five batches of documents (9-13 documents in each batch), and each batch was independently annotated by two different annotators. After the completion of the first batch (*development set*, 10 documents), we jointly discussed issues that had arisen during the annotation process, to refine and reach a consensus on improvements and edits in the guidelines. As a result, we added specific rules for the time expressions that had caused disagreements, and removed ambiguous sentences and examples. For instance, we found that "dates" and "durations" were sometimes hard to distinguish, and created specific rules to disambiguate those

[4] PrePostExp and Quantifier TIMEX types were not considered.

(e.g., *over the last week* should be annotated as a Duration, and not a Date). The updated guidelines were then applied to annotate the remaining documents. When all batches had been double-annotated, we carried out an adjudication phase in order to create a gold standard corpus. The adjudicator decided which annotations to include in the gold standard in case of disagreement between annotators, added missing annotations and omitted or corrected erroneous ones.

2.3 Automated Time Expression Extraction

In this study, we explored two well-known time expression taggers, SUTime (Chang and Manning, 2012) and HeidelTime (Strötgen and Gertz, 2010), which were developed and evaluated on general domain corpora. When applied to the TempEval-2 newspaper data, both systems achieved state-of-the-art performance (F1 scores of 92% and 86%, respectively, for time expression identification) (Verhagen et al., 2010). Moreover, they have previously been used for the automatic processing of clinical narratives (Jindal and Roth, 2013; Wang et al., 2016).

Both SUTime and HeidelTime use a list of pattern matching rules, built on regular expressions, to recognize and normalize time expressions inside the text. As a main difference, while SUTime links relative TIMEXes (such as *yesterday*) to the document creation date, HeidelTime uses different normalization strategies depending on documents' types (e.g., *news*, *narratives*).

To adapt the systems to the mental health domain, we first evaluated their original versions on the development set[5], to see what the increase in performance over non-domain-specific rules would be. Then, we manually reviewed the TIMEXes present in the development set, and modified and added rules as needed. The performance of the updated systems was then evaluated on a *validation set*, consisting of two batches (23 documents in total). To allow for future development and evaluation, we did not use the remaining batches (*test set*, 19 documents) in this study. The documents we used to adapt and evaluate the temporal taggers were from the adjudicated gold standard corpus.

[5]To compute the performance of the original systems, we used: the SUTime grammar included in the Stanford CoreNLP (*https://stanfordnlp.github.io/CoreNLP/*) distribution dated 2017-06-09, and the HeidelTime resources included in the GATE (*https://gate.ac.uk/*) distribution 8.3.

2.4 Evaluation Metrics

To assess the quality of the developed corpus, we calculated inter-annotator agreement (IAA) for each annotated batch, using the metrics that were used for i2b2 2012 (average of precision and recall) and Clinical TempEval 2016 (F1 score). First, we computed the average of precision and recall: the entities marked by one annotator were used as the gold reference, while the entities identified by the second annotator were considered as the system's output (switching these two roles would not change the final result). Moreover, we measured the F1 score (i.e., the harmonic mean of precision and recall), which provides a good way to quantify agreement for entity extraction tasks (Hripcsak and Rothschild, 2005).

To evaluate the performance of SUTime and HeidelTime, we defined: (i) true positives (TP), as the gold TIMEXes that were found in the system's output; (ii) false negatives (FN), as the gold TIMEXes that were not found in the system's output; and (iii) false positives (FP), as the system TIMEXes that were not found among gold annotations. In this case, we assessed the system's performance in terms of precision, recall, and F1 score.

3 Results

3.1 Time Expression Annotations

The total number of gold TIMEXes in our corpus is 3,413, with an average of 65.6 annotations per document[6]. Table 1 reports the prevalence of TIMEX types in the corpus, divided into development, validation, and test sets. Overall, the majority of TIMEXes are represented by Dates (55.8%). Durations, Times, and Frequencies account for 16.5%, 10.7%, and 8.1%, respectively.

As mentioned, we defined a new TIMEX type referring to the patient's age: "Age-related". This type was assigned to 8.9% of all TIMEXes. Some examples include:

- *at the age of 8*: requires adding 8 years to the date of birth for normalization;
- *when he was a child*: requires the date of birth and a shared definition of "child years" for normalization;

[6]Annotators worked 20-24 hours, and annotated 2/3 of the corpus each (33-39 docs): the average time required for corpus development was around 35-40 minutes per document.

	development set	validation set	test set	total
# documents	10	23	19	52
# TIMEXes	964 (96.4/doc)	1,401 (60.9/doc)	1,048 (55.2/doc)	3,413 (65.6/doc)
Date	593 (61.5%)	803 (57.3%)	507 (48.4%)	1,903 (55.8%)
Duration	148 (15.3%)	215 (15.3%)	200 (19.1%)	563 (16.5%)
Time	94 (9.8%)	129 (9.2%)	143 (13.6%)	366 (10.7%)
Frequency	60 (6.2%)	127 (9.1%)	89 (8.5%)	276 (8.1%)
Age-related	69 (7.2%)	127 (9.1%)	109 (10.4%)	305 (8.9%)

Table 1: TIMEX annotation results: prevalence of types in our corpus.

- *since his teens*: requires the date of birth and a shared definition of "teens years" for normalization;
- *during his first year* (implicitly referring to the first year of university): requires the date of birth and the usual timing of university for normalization;

With respect to IAA, we computed results on TIMEX spans (without considering the different TIMEX types, as this was not calculated for the corpora used for comparison), for both partial and exact matches. In the first case, the average of precision and recall was 78%, and the F1 score was 77%. In the second case, both metrics resulted in 60%.

For partial matches, the IAA per batch was in the range of 73.6%-83.7% (average of precision and recall), and 73.5%-83.3% (F1 score). We also computed the percentage of TIMEX type matches for those time expressions that the annotators agreed on with respect to overlapping spans, resulting in 91% percentage match.

3.2 Comparison to Related Corpora

In Table 2, our corpus is compared to the i2b2 2012 and the Clinical TempEval 2016 corpora. Specifically, the table reports the size, the number of TIMEXes, the type prevalence[7], and the IAA values for the three considered corpora. To allow comparing TIMEX types among these corpora, we merged Clinical TempEval annotations as follows: PrePostExp time expressions were considered among Dates, while Quantifier time expressions were considered as Frequencies. No modifications were required in order to compare the i2b2 2012 corpus. Also, since we added the new TIMEX type Age_related, we were not able to compare these annotations in either corpus.

3.3 Temporal Expression Extraction System Adaptation

In this work, we used SUTime and HeidelTime to identify TIMEX spans in the developed corpus[8]. The results of this domain adaptation are shown in Table 3. First, we ran the original versions of the two systems on the development set, obtaining an F1 score of 72.5% for SUTime and 63.6% for HeidelTime (allowing partial matches). As expected, these scores are much lower than those obtained on general domain corpora (92% and 86% F1 scores on TempEval-2 newspaper data). After tuning the systems' rules on the development set, we achieved scores of 79.7% and 77.3%, respectively. By running the updated systems on the validation set, we obtained a final result of 79.5% and 75.8%, respectively.

It is important to mention that, although the original version of SUTime included rules to capture some "age" expressions (e.g., *28-year-old*), these were considered as Durations. In the original version of HeidelTime, instead, these expressions were explicitly excluded, as they were probably not considered as proper time expressions. This is one of the reasons why the original version of HeidelTime had much lower recall than SUTime (Table 3, "HeidelTime original" row).

4 Discussion

Extracting temporal information from mental health records is particularly challenging, as this domain is characterized by a large proportion of free-text and heterogeneity in self-reported experiences (i.e., mental health symptoms), circum-

[7]These numbers were computed on released data, for i2b2 2012, and on publicly available annotations, for Clinical TempEval 2016.

[8]For determining Age-related TIMEXes, we applied a set of post-processing rules to the output of the two temporal taggers.

	Our corpus	**i2b2 2012**	**Clinical TempEval 2016**
Domain	Mental health	Intensive care	Oncology
# documents	52	310	591
# tokens	206,661 (3,974/doc)	178,000 (574/doc)	550,487 (931/doc)
# TIMEXes	3,413 (1.65/100tok)	4,184 (2.35/100tok)	7,863 (1.43/100tok)
Prevalence	Date: 55.8% Duration: 16.5% Time: 10.7% Frequency: 8.1% Age_related: 8.9%	Date: 68.4% Duration: 17.8% Time: 3.1% Frequency: 10.7% Age_related: NA	Date: 76.1% Duration: 10.6% Time: 3.4% Frequency: 9.9% Age_related: NA
IAA (Avg P-R)	Partial: 78% Strict: 60%	Partial: 89% Strict: 73%	NA
IAA (F1 score)	Partial : 77% Strict : 60%	NA	Partial: NA Strict: 73%

Table 2: Comparison between our corpus, i2b2 2012, and Clinical TempEval 2016. IAA: inter-annotator agreement; Avg P-R: average of precision and recall; NA: not applicable (TIMEX type not annotated or IAA metric not calculated in these corpora).

Set	**System**	**P**	**R**	**F1**
dev	SUTime original	71.4%	73.6%	72.5%
	HeidelTime original	71.7%	57.2%	63.6%
dev	SUTime updated	72.9%	87.8%	79.7%
	HeidelTime updated	73.6%	81.3%	77.3%
valid	SUTime updated	72.8%	87.7%	79.5%
	HeidelTime updated	70.5%	81.9%	75.8%

Table 3: SUTime and HeidelTime results. P: precision; R: recall.

stances (e.g., social support networks, recent or past stressful experiences, psychoactive substance use), and treatment and outcomes. In this study, we annotated time expressions related to patients with schizophrenia in EHRs. The documents in our corpus are long when compared to similar corpora (3,974 tokens/doc), and include a large proportion of relevant time expressions (65.6 TIMEXes/doc). In addition, they might contain information taken from structured forms (e.g., questions, references to health care legislation), which are not relevant to the patient's clinical history, but could still include references to time.

4.1 Comparison to Related Corpora

When comparing our corpus to other related corpora, there are differences in the documentation types that can have an impact on the development of temporal information extraction systems. For instance, the discharge summaries in the 2012 i2b2 corpus each start with the admission and discharge date, which are annotated as TIMEXes. Similarly, the Clinical TempEval 2016 documents are organized in sections with semi-structured date information, that can be useful to then link and anchor clinically relevant events in the documents. The documents in our corpus, on the contrary, include various paragraphs describing both past and current events related to the patient, without necessarily following a predefined structure.

Regarding TIMEX types, there was a greater prevalence of Date expressions in the i2b2 2012 (ICU domain) and Clinical TempEval 2016 (oncology domain) corpora, as compared to our corpus (Table 2). This might relate to the fact that, in the ICU and oncology clinical settings, treatment episodes are likely to be shorter and changes in physical health parameters and onset/duration of treatment occur over shorter periods of time. As another interesting observation, our corpus is characterized by a higher prevalence of the Time type, which is probably due to the fact that many events are described as happening at a specific time of day (*this morning*, *at night*). It is important

to point out that in both i2b2 2012 and Clinical TempEval 2016, age-related information was not marked. One reason for this might be that these types of constructs were not considered useful for the use-cases that these corpora were developed for.

As regards the IAA, we obtained a value of 60%/78% (strict/partial) for the average of precision and recall, and a value of 60%/77% (strict/partial) for the F1 score. Although these results are lower in comparison to those of i2b2 2012 and Clinical TempEval 2016 (Table 2), this can be considered a promising result, given the intrinsic complexity of the mental health domain. As an important remark, the difference between partial and strict IAA measures indicates that identifying the spans of time expressions is not straightforward. This is also reflected in the results obtained on the i2b2 2012 corpus. In our case, the main reason for disagreement was the inclusion/exclusion of prepositions or determiners in TIMEX spans (e.g., *for three years* instead of *three years*). We also analyzed disagreements in TIMEX type assignments. Differentiating between Date and Duration was one of the main disagreements (accounting for 42% of disagreements). For instance, an expression like *this week* was assigned Date (interpreted as a point in time) by one annotator, and Duration (interpreted as a period) by another.

4.2 Domain-specific Time Expressions

As an interesting result of the annotation task, we identified a set of time expressions which were not present in the other corpora, but which are essential to allow capturing symptom onset. As previously mentioned, these expressions are those related to the age of the patient, which account for 8.9% of all TIMEXes (Table 1). Despite not being particularly frequent, Age-related TIMEXes can be crucial to determine the first onset of symptoms, which is often reported by patients or their relatives in a vague way. For example, extracting these kinds of expressions is essential for sentences like[9]:

- *she first experienced hallucinations* ***at the age of 18***
- *he started hearing voices* ***when he was 15***
- *he has been experiencing these symptoms* ***since his teens***

[9]The reported sentences have been paraphrased.

Besides defining a new TIMEX type, we also found some example TIMEXes that are specific to the analyzed domain and were not present in the compared corpora. As a first example, we identified a few expressions that are related to drug prescriptions, such as *OD* (i.e., once daily) and *NOCTE* (i.e., every night). Moreover, we noticed that the expression *15 minute* is often used as a Frequency, rather than a Duration, as this is the usual interval of time used to observe patients with schizophrenia (e.g., "*he was placed on* ***15 minute*** *visual observations*"). Determining the correct interpretation is not straightforward, as this relies on domain knowledge (in the sentence *I went for a* ***15-minute*** *walk*, the same TIMEX represents a Duration). As another interesting example, we realized that, in the field of mental health, the expressions */7*, */12*, and */52* can be used to refer to days (*3/7 ago* = three days ago), months (*for 3/12* = for three months), and weeks (*in 2/52* = in two weeks), respectively. In our corpus, we found a total of 12 expressions of this kind (4 Dates, 6 Durations, and 2 Age_related). To normalize them, it is possible to create specific rules that map the different patterns to the corresponding temporal values.

4.3 Time Expression System Adaptation

We applied SUTime and HeidelTime on the development set through an iterative process (Table 3). By running the two original versions of the systems, we found that SUTime performed better than HeidelTime, especially in terms of recall (73.6% versus 57.2%): this is probably due to the fact that, in our annotation schema, we included a few expressions which were already taken into account by the first system, but not by the second (e.g., *28 years old*, *past*). In the adaptation process, we identified false negatives (FNs) for both systems, and then refined rules to capture them. It is important to point out that, in this preliminary experiment, we focused on improving the systems' performance for partial matches, rather than identifying exact TIMEX spans. While a few of the performed adaptations involved general domain rules (e.g., dates in the form "dd/MM" were not recognized by SUTime), we mostly needed to address TIMEXes specific to the mental health domain. By adding extraction rules for these expressions, we were able to reduce the number of FNs, thus obtaining an improvement in recall from 73.6%

to 87.8%, for SUTime, and from 57.2% to 81.3% for HeidelTime (development set). As for precision, lowering the number of FPs was not trivial, as these rule-based systems cannot distinguish between time expressions that are patient-related (thus relevant to our goal) and those that are not (e.g., form-related).

After an error analysis, we found a few non-trivial TIMEXes that were not correctly captured by the system and will require further adaptation/improvement. For instance, SUTime was not able to identify age-related expressions like *between the ages of 10 and 12* and Time intervals like *9.30-10*. On the other hand, ambiguous words such as *present* (as in *present at the appointment*) and *minutes* (as in *minutes of the meeting*) were erroneously considered as TIMEXes. Also, all the time expressions that were included in form-like paragraphs (e.g., *The Activities of **Daily** Living include...*) were counted among false positives, as we were not interested in extracting these.

In this study, the best final F1 score was obtained with the adapted version of SUTime (79.5% on the validation set), which represents a promising result if compared to the IAA of 77% (F1 score). This could reflect the fact that time expressions often follow specific patterns: by adequately tuning extraction rules, it is possible to obtain a good extraction performance, which can be even higher than that of a human annotator (this is particularly true for recall).

4.4 Limitations and Future Work

As a main limitation of this work, we only addressed the extraction of TIMEX spans and types, without dealing with TIMEX value normalization, which would require assigning a standardized value to each TIMEX. For instance, the expression *for three years* should be normalized as *P3Y*, while the expression *at the age of 8* would require the date of birth in order to be correctly normalized. We are in the process of extending our TIMEX annotations with normalized values. In future work, we will use these annotations to develop suitable rules for time expression extraction and value normalization. Moreover, while adapting TIMEX extraction systems, we did not write contextual rules to disambiguate expressions that can belong to different types depending on their context, although we did disambiguate these during manual annotation (e.g., *at night* was marked as a Time, when referring to a single episode, or as a Frequency, when referring to a drug prescription). As a future improvement, we will address this task by dealing with the context in which time expressions appear, for example, by using word embeddings to represent each word with automatically derived contextual features (Mikolov et al., 2013). Finally, since we are interested in assessing the usability of SUTime and HeidelTime in other clinical domains, we plan to extend the adaptability study presented in this work to other clinical corpora, such as the i2b2 and Clinical TempEval corpora used for comparison here.

As previously mentioned, the extraction of time expressions represents a first step towards our final goal, i.e., the identification of symptom onset and DUP in free text. The next step will involve the annotation of clinically relevant entities (symptoms primarily), to be linked to the available temporal information. Extracting entities such as symptoms could be done by knowledge-based approaches based on keyword lists, or using word embedding models (or a combination of both). We are currently exploring different alternatives. As for temporal linking, we will need to refer each clinical entity to a specific time expression. For example, given the sentence *he first experienced hallucinations in 2008*, the following link should be identified: "{2008} CONTAINS {hallucinations}". To reach this goal, we are currently experimenting with the annotation of a set of documents, where relevant events and time expressions have been pre-annotated by using automatic approaches.

5 Conclusion

In this paper, we described the annotation of time expressions in mental health records related to schizophrenia, thus creating an annotated corpus. To the best of our knowledge, this is the first gold standard developed in this domain for a specific mental health use-case: onset and DUP extraction. In addition, this is the first study explicitly incorporating age-related information, which is not captured by current temporal models. As an important aspect, we also assessed the adaptability of two existing rule-based TIMEX extraction systems to the new analyzed use-case, obtaining promising results.

Due to governance regulations, the corpus annotated in this study cannot be made publicly available. However, there are procedures in place

to provide researchers with controlled access to the CRIS database. Moreover, the developed guidelines and the adapted versions of SUTime and HeidelTime have been made publicly available[10], and could be easily reused or adapted for other temporal information extraction tasks.

Acknowledgments

NV and SV are supported by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund, Project INCA 600398.

RS, RD and RP are funded by the NIHR Specialist Biomedical Research Centre for Mental Health at the South London and Maudsley NHS Foundation Trust and Institute of Psychiatry, King's College London.

RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship (MR/S003118/1) and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK.

References

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

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.

Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *LREC*.

Charlotte Connor, Max Birchwood, Colin Palmer, Sunita Channa, Nick Freemantle, Helen Lester, Paul Patterson, and Swaran Singh. 2013. Don't turn your back on the symptoms of psychosis: a proof-of-principle, quasi-experimental public health trial to reduce the duration of untreated psychosis in Birmingham, UK. *BMC psychiatry*, 13:67.

[10]The guidelines developed in this study are available at: *https://github.com/medesto/annotation-guidelines*. The adapted versions of SUTime and HeidelTime are available at: *https://github.com/medesto/systems-adaptation*.

Michele Hill, Niall Crumlish, Mary Clarke, Peter Whitty, Elizabeth Owens, Laoise Renwick, Stephen Browne, Eric A. Macklin, Anthony Kinsella, Conall Larkin, John L. Waddington, and Eadbhard O'Callaghan. 2012. Prospective relationship of duration of untreated psychosis to psychopathology and functional outcome over 12 years. *Schizophrenia research*, 141(2-3):215–221.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

Prateek Jindal and Dan Roth. 2013. Extraction of events and temporal expressions from clinical narratives. *Journal of biomedical informatics*, 46:S13–S19.

Stephen Kisely, Anita Scott, Jennifer Denney, and Gregory Simon. 2006. Duration of untreated symptoms in common mental disorders: association with outcomes. *The British Journal of Psychiatry*, 189(1):79–80.

Julia M. Lappin, Kevin D. Morgan, Craig Morgan, Paola Dazzan, Abraham Reichenberg, Jolanta W. Zanelli, Paul Fearon, Peter B. Jones, Tuhina Lloyd, Jane Tarrant, Annette Farrant, Julian Leff, and Robin M. Murray. 2007. Duration of untreated psychosis and neuropsychological function in first episode psychosis. *Schizophrenia research*, 95(1-3):103–110.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–144.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Gayan Perera, Matthew Broadbent, Felicity Callard, Chin-Kuo Chang, Johnny Downs, Rina Dutta, Andrea Fernandes, Richard D Hayes, Max Henderson, Richard Jackson, Amelia Jewell, Giouliana Kadra, Ryan Little, Megan Pritchard, Hitesh Shetty, Alex Tulloch, and Robert Stewart. 2016. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open*, 6(3).

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

Brett R South, Shuying Shen, Jianwei Leng, Tyler B Forbush, Scott L DuVall, and Wendy W Chapman. 2012. A prototype tool set to support machine-assisted annotation. In *Proceedings of the 2012*

Workshop on Biomedical Natural Language Processing, pages 130–139.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324.

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.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62.

Wei Wang, Kory Kreimeyer, Emily Jane Woo, Robert Ball, Matthew Foster, Abhishek Pandey, John Scott, and Taxiarchis Botsis. 2016. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. *Journal of biomedical informatics*, 62:78–89.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of biomedical informatics*, 77:34–49.