

Rafael Faria de Azevedo

**Temporal Tagging of Noisy Clinical Texts Written in
Brazilian Portuguese**

Dissertation presented to the Graduate Program in Informatics of the Pontifícia Universidade Católica do Paraná (PUCPR) as a partial requirement for the degree of Master in Informatics.

Curitiba

2019

Rafael Faria de Azevedo

**Temporal Tagging of Noisy Clinical Texts Written in
Brazilian Portuguese**

Dissertation presented to the Graduate Program in Informatics of the Pontifícia Universidade Católica do Paraná (PUCPR) as a partial requirement for the degree of Master in Informatics.

Major Field: Computer Science

Advisor: Emerson Cabrera Paraiso

Curitiba

2019

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Edilene de Oliveira dos Santos CRB 9 /1636

A994t
2019 Azevedo, Rafael Faria de
Temporal tagging of noisy clinical texts written in brazilian portuguese /
Rafael Faria de Azevedo ; advisor, Emerson Cabrera Paraiso. -- 2019
149 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná,
Curitiba, 2019
Bibliografia: f.143-149

1. Processamento eletrônico de dados. 2. Software – Desenvolvimento. 3.
Cuidados médicos. 4. Tecnologia médica. 5. Tabela de temporalidade. I.
Paraiso, Emerson Cabrera. II. Pontifícia Universidade Católica do Paraná.
Programa de Pós-Graduação em Informática. III. Título.

CDD 20. ed. – 004

DECLARAÇÃO

Declaro para os devidos fins que o aluno **RAFAEL FARIA DE AZEVEDO**, defendeu sua dissertação de Mestrado intitulada “**TEMPORAL TAGGING OF NOISY CLINICAL TEXTS WRITTEN IN BRAZILIAN PORTUGUESE**”, na área de concentração Ciência da Computação, no dia 16 de abril de 2019, no qual foi aprovado.

Declaro ainda que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade, firmo a presente declaração.

Curitiba, 03 de julho de 2019.



Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná

Acknowledgements

First of all I would like to thank my advisor, Emerson Cabrera Paraiso, for his full support in this journey of knowledge. I thank my family, especially my parents, Matilde Faria de Azevedo (in memory) and Nelson Moreira de Azevedo. I am grateful to my sister Juliana Faria de Azevedo and my brother Matheus Palasi de Azevedo, as well as my stepmother Luciana Cristina Palasi de Azevedo. I also thank Professor Claudia Maria Cabral Moro Barra and PUCPR students Adalniza Moura Pucca da Silva, Bruno Henrique da Silva Santos, Caroline Pilatti GebelUCA, Letícia Kist Mantovani and Mayara Regina da Silva Reis for participating in the construction of the corpus used in this work. I am also grateful to all teachers, staff and colleagues of PPGIa, especially Zacarias Curi Filho, Cheila Cristina Farias, Luis Eduardo Boiko Ferreira, Jhonatan Geremias, Lucas Martiniano de Oliveira, Rodolfo Miranda Pereira, João Pedro Santos Rodrigues, Flávia Beuting and Sandoval Ribas Ruppel. I worked at Banco do Brasil for the entire period of the master's degree, so I thank all workmates who collaborated to make this project successful. Special thanks to Kalinka Trzaskos, Vanessa Cristiane Motta de Matos, Emmanuelle Regiane Cunha de Oliveira, Roberto Vitor Sabela Pereira and Klailton Ralff da Silva. I am also grateful to the support of my friends and neighbours Eloy Klaus Quinhones and Eder Jofre Cleomar Quinhones. Finally, I would like to thank the Brazilian Government, once this study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

*“The important thing is not what they do to us,
but what we ourselves do of what they have done to us”*

*“O importante não é o que fazem de nós,
mas o que nós mesmos fazemos do que fizeram de nós”*

Jean Paul Sartre

Abstract

Temporal tagging in clinical texts allows gathering useful information for different kind of healthcare software. It is important in automatic tasks such as, patient treatment report summarization, timeline elaboration drawing on chronic disease patient clinical reports, or querying temporal information in clinical narratives. With the massive increase of Electronic Health Records (EHR) in hospitals and clinics, recognition and normalization of temporal expressions has become an important input to organize chronologically reports that record patients' ongoing evolution during their treatment. Insights provided by this kind of text processing give physicians and nurses a better understanding about patients conditions, especially those undergoing chronic treatment. Historically, systems that deal with temporal tagging in clinical texts were made to work with English well-written texts, most of them designed to cope with data from temporal information challenges like the Informatics for Integrating Biology and the Bedside (i2b2), the Clinical TempEval task, within the Semantic Evaluation (SemEval) challenge and others. Clinical texts may be noisy in some circumstances. In this study, the word noisy indicates any type of text produced in a non-regular pattern of language or non standard way, like misspelling, typing errors, a forgotten or duplicated character etc. The aim of this study is to deal with temporal expression tagging in noisy clinical texts written in Brazilian Portuguese, an under-resourced language in the clinical domain as well as in temporal tagging tasks. To do that, new rules were added to a rule-based temporal tagger called HeidelTime. The new rules were drawn from both well-written temporal expressions and noisy ones based on texts compiled from Brazilian hospital sources. Three different strategies were used to tag and normalize as many temporal expressions as possible in the compiled noisy clinical texts. A first strategy relied on using correctly written temporal expressions found in the corpus retrieved from clinical reports. A second strategy used noisy temporal expressions found in the same corpus. Finally, a third strategy used an n-gram approach to improve the extraction and normalization of not well-written temporal expressions from the same clinical reports. In order to accomplish the planned strategies, it was necessary to annotate the aforementioned clinical corpus, which is also a contribution of this work. The annotation process was done from scratch by health specialists and is detailed in this document. The different approaches results were compared and the best result reached a F1-Score of 0.9021 in the extraction step and a F1-Score of 0.9245 in the normalization step. Our results are on the order of other works done with the same aim, considering the Brazilian Portuguese and other languages, whether coping with noisy temporal expressions or not.

Key-words: temporal tagging, temporal expression, noisy clinical report, real world clinical text, rule-based system.

List of Figures

Figure 1 – Temporal tagging over a clinical report written in Portuguese	15
Figure 2 – Temporal tagging over a noisy clinical report written in Portuguese . . .	16
Figure 3 – Extraction and normalization of temporal expressions (Extracted from Strötgen and Gertz (2016)).	23
Figure 4 – The main types of well-defined relationships between two points in time (Extracted from Strötgen and Gertz (2016)).	23
Figure 5 – Temporal information normalization process (Extracted from Strötgen and Gertz (2016)).	24
Figure 6 – Temporal information ordered hierarchically (Extracted from Strötgen and Gertz (2016)).	25
Figure 7 – Four realization types of temporal expressions (Extracted from Strötgen and Gertz (2016)).	27
Figure 8 – Precision formula	32
Figure 9 – Recall formula	32
Figure 10 – F1-measure formula	33
Figure 11 – Accuracy formula	33
Figure 12 – Normalization-accuracy formula	33
Figure 13 – HeidelTime architecture with algorithm (source code) and resources. Extracted from Strötgen and Gertz (2013).	35
Figure 14 – Pattern resource files for different expressions for month names and numbers (a) and normalization resource file for month expressions (b). (c) represents how their information is used in HeidelTime after being read and translated by HeidelTime’s resource interpreter. Extracted from Strötgen and Gertz (2013).	35
Figure 15 – HeidelTime’s algorithm overview. Extracted from Strötgen and Gertz (2013).	41
Figure 16 – Research structure developed in this study	51
Figure 17 – Method processing step	54
Figure 18 – Example of created rules and a normalization resource	56
Figure 19 – Files split strategy	59
Figure 20 – Files split by clinical text character quantity	60
Figure 21 – Pre-annotation samples	61
Figure 22 – Screenshot of the MAE tool custom version used by annotators	63

Figure 23 –Screenshot of the MAE tool version used by the adjudicator	64
Figure 24 –Approximate percentage of TEs type in the created corpus	69

List of Tables

Table 1 – TIMEX2 attributes (Extracted from Ferro et al. (2005))	28
Table 2 – Values for MOD attribute (Extracted from Group et al. (2009))	30
Table 3 – The setup of a confusion matrix, used to evaluate temporal taggers (Ex- tracted from Strötgen and Gertz (2016))	32
Table 4 – Line break problem	53
Table 5 – Distribution of files between annotator pairs	61
Table 6 – Unambiguous examples of SET type	71
Table 7 – Confusing and ambiguous examples of SET type	72
Table 8 – Quantity of altered rules in all experiments	74
Table 9 – Quantity of altered rules in all experiments	74
Table 10 –Quantity of TP TEs extracted by HeidelTime which were used to eval- uate the normalization step	84
Table 11 –Sample standard deviation of extraction and normalization steps made by HeidelTime in each experiment	84
Table 12 –Extraction experiments results	85
Table 13 –Normalization experiments results	85
Table 14 –Significance of Wilcoxon Test in extraction	86
Table 15 –Significance of Friedman and Dunn-Bonferroni tests in extraction for all experiments	86
Table 16 –Significance of Friedman and Dunn-Bonferroni tests in extraction for 3 experiments	87
Table 17 –Significance of Wilcoxon Test in normalization	87
Table 18 –Significance of Friedman and Dunn-Bonferroni tests in normalization for all experiments	88
Table 19 –Significance of Friedman and Dunn-Bonferroni tests in normalization for 3 experiments	88
Table 20 –Abreviações	124
Table 21 –Valores possíveis do atributo mod	127
Table 22 –Abbreviation table	136
Table 23 –Mod options	139

List of abbreviations and acronyms

BIO	Begging-Inside-Outside
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
EHR	Electronic Health Record
FN	False Negative
FP	False Positive
IR	Information Retrieval
ISO	International Organization for Standardization
LSTM	Long Short-Term Memory
NER	Named Entity Recognition
NLP	Natural Language Processing
PT	Period Time
Regex	Regular Expression
SMS	Short Message Service
SQL	Structured Query Language
SVM	Support Vector Machine
TE	Temporal Expression
TER	Temporal Expression Recognition
TEs	Temporal Expressions
Timex	Temporal Expression
Timexes	Temporal Expressions
TN	True Negative
TP	True Positive
TPZ	Time Point Zero
val	Value

Contents

1	Introduction	14
1.1	Motivation	17
1.2	Objectives	20
1.3	Hypothesis of the work	20
1.4	Scientific Contribution	21
1.5	Scope	21
1.6	Document Organization	21
2	Basic concepts	22
2.1	The word "noisy"	22
2.2	Temporal Tagging	22
2.3	Metrics	31
2.4	HeidelTime	34
2.4.1	HeidelTime's architecture	34
2.4.2	HeidelTime's resources	35
2.4.3	HeidelTime's rule syntax	36
2.4.4	HeidelTime's algorithm	41
3	State-of-the-art	42
3.1	Annotated corpora	43
3.2	Rule-based approach	45
3.3	Machine learning and hybrid approach	46
3.4	SemEval 2015, 2016 and 2017	48
4	Methodological procedures	50
5	Proposed method	52
5.1	Method assumptions	53
5.2	Preprocessing	53
5.3	Processing	53
6	Dataset annotation	58
6.1	The dataset	58
6.2	The annotation process	59
6.2.1	The annotators and the adjudicator	59
6.2.2	The dataset preparation	59
6.2.3	The annotation	62

6.2.3.1	Annotators training	62
6.2.3.2	Final annotation	64
6.2.3.3	Adjudication	64
6.3	The guideline	72
7	Experiments and results	73
8	Discussion	89
9	Conclusion	117
	Appendix	119
APPENDIX A	Guideline	120
A.1	Motivação	120
A.2	Expressão Temporal	120
A.3	Extração	120
A.4	Normalização	121
A.5	Orientações importantes sobre como fazer a normalização das expressões temporais	121
A.5.1	Idioma para normalização	121
A.5.2	ET ruidosas	121
A.5.3	A justa medida de uma ET	122
A.6	Atributo type	122
A.7	Categorias de expressões temporais	124
A.8	Atributo value (val)	126
A.9	Atributo mod	126
A.10	O atributo freq	128
A.11	O que não marcar?	128
A.12	Protocolo de anotação	129
APPENDIX B	Guideline	132
B.1	Motivation	132
B.2	Temporal expression	132
B.3	Extraction	132
B.4	Normalization	133
B.5	Important guidelines on how to normalize temporal expressions	133
B.5.1	Language for normalization	133
B.5.2	Noisy temporal expression	133
B.5.3	The fair measure of a TE	133

B.6	The type attribute	134
B.7	Categories of temporal expressions	136
B.8	The value (val) attribute	138
B.9	The mod attribute	138
B.10	The freq attribute	140
B.11	What not to tag?	140
B.12	Annotation protocol	141
Bibliography	144

1 Introduction

Time plays a central role in any information space. It is quite common to find words which indicate time in most of text types. To take advantage of such time related data, it is important to extract it, anchor it in time, compute it, and make all this temporal information available to subsequent exploration tasks (ALONSO et al., 2011). Temporal information retrieval has been a topic of interest in recent years. The scope of this interest is to improve the effectiveness of this information retrieval by different types of methods (CAMPOS et al., 2015). Retrieval of temporal information from text is important to many text processing tasks which require deep language understanding (MENG; RUMSHISKY; ROMANOV, 2017) within Natural Language Processing (NLP). This can be useful in many areas such as question answering, search, text classification, in systems that establish timelines and in text summarization (SARATH; MANIKANDAN; NIWA, 2017), (UZZAMAN; ALLEN, 2010), (SCHILDER, 2004).

From a conceptual point of view, there are three tasks involving time and its related entities. The first one is temporal tagging, which comprises two steps, temporal expression recognition (extraction) and normalization (KOLOMIYETS; MOENS, 2010), (ROBERTS; RINK; HARABAGIU, 2013). The second task is event tagging. The third task is temporal relation. Temporal tagging is the process of finding a word or a set of words that indicate time in a given text, also called a **temporal expression** (TE). Event tagging is the process of finding a word or a group of words that indicate an event in a given text. Temporal relation is the link between a TE and its related event. The whole process, which groups temporal tagging, event tagging and temporal relation, is called full temporal annotation task (VERHAGEN et al., 2009) or temporal information extraction (STRÖTGEN; GERTZ, 2016).

This work is concerned with temporal tagging tasks. They comprise (i) a temporal expression recognition (TER) step, which is the process of finding a word, a set of words or a mix of words, numbers and some special characters that is a TE; and (ii) a normalization step (performed after TER). Normalization is the process of transforming a TE into a standard that can be understood by humans and other systems. The broad general standard used in the temporal information research area is established by an International Organization for Standardization¹ (ISO) is called TimeML (PUSTEJOVSKY et al., 2003), (PUSTEJOVSKY et al., 2005).

¹ <https://www.iso.org/home.html>

xxxxxx,
 <TIMEX3 value="P82Y" type="DURATION" tid="t1">82 anos</TIMEX3>
 #has com tratamento irregular #dm2- amaurose a d #tabagismo passivo #fibrilação atrial com
 tempo de início indeterminado paciente veio à emergência dia 24.05 com dor anginosa típica ja
 com
 <TIMEX3 value="PT11H" type="DURATION" tid="t5">11 horas</TIMEX3>
 de evolução e com pa de 178/122mm hg. tinha alteração de ecg (supra desnivelamento de st de
 v1 a v3, com inversão de t difusa). foi interpretado inicialmente como sca com ssst e levado a
 cateterismo com
 <TIMEX3 value="PT15H" type="DURATION" tid="t6">15 horas</TIMEX3>
 desde o início da dor. o exame mostrou 2 lesões graves em acd sendo feita angioplastia com
 implante de 2 stents. a paciente evoluiu bem clinicamente e após o resultado do exame (pela
 discordância de localização entre o ecg e cateterismo) os sintomas e elevação de marcadores
 cardíacos foram atribuídos à crise hipertensiva. com
 <TIMEX3 value="PT48H" type="DURATION" tid="t7">48 horas</TIMEX3>
 do procedimento houve formação de hematoma volumoso no local da punção, sem
 pseudoaneurisma ou sangramento ativo visto em eco doppler de mid. avaliado pela cirurgia
 vascular que indicou tratamento conservador por considerar o hematoma em regressão. foi
 verificado também infecção do trato urinário, com tratamento por
 <TIMEX3 value="P3D" type="DURATION" tid="t2">3 dias</TIMEX3>
 (1d de norfloxacin e 2d de cefuroxime). recebe alta em boas condições clínicas para
 acompanhamento no ambulatório da cardiologia em pós-alta, quando deverá ser decidido por
 anticoagulação (momento atual inadequado pelo hematoma pós-procedimento). #recomendações
 pós-alta# -retorno à emergência em caso de qualquer intercorrência -uso das medicações
 conforme prescrito: aas 100mg, clopidogrel 75 mg por
 <TIMEX3 value="P2D" type="DURATION" tid="t3">2 dias</TIMEX3>
 e ticlopidina 250mg 2x por
 <TIMEX3 value="P1M" type="DURATION" tid="t4">1 mes</TIMEX3>
 , sinvastatina 40mg 1x, enalapril 20mg 2x, metoprolol 50mg 2x, isossorbida se necessario,

Figure 1 – Temporal tagging over a clinical report written in Portuguese

Figure 1 shows an example of temporal tagging. The word or a set of words surrounded by the tag TIMEX3 indicate time; this is a TE. The content of the attribute “value” in the tag indicates the marked TE normalized. The content of the tag attribute “type” indicates the type of the TE; more details about this process are presented in chapter 2.

Temporal tagging can be performed on different types of text such as news reports, narratives, short messages, chats and scientific articles. Clinical text is also an important input to obtain this kind of information. The increasing adoption of electronic health records (EHR) has enabled health institutions to gather a huge amount of data about their patients. However, this amount of data is growing up fast, and what was initially seen as an advance ended up turning into a hard challenge, data needs to be processed and transformed into valuable information. In order to organize this data, temporal tagging plays an important role, because it is the first step towards other tasks related to temporal information within the clinical domain such as: summarizing patient treatment, timeline development to follow patient evolution during treatment, search for information related to time etc. (PIVOVAROV; ELHADAD, 2015).

Processing temporal information in clinical text is more difficult than in news reports (the most studied domain in the temporal information research area) due to implicit expression of temporal information, domain specificity, lack of structure and writing qual-

ity (MOHARASAR; HO, 2016). Figure 2 illustrates this aspect of clinical text, which is noisy many times.

In this study, the word noisy indicates any type of text which is a temporal expression (TE) produced in a non-regular pattern of language or non-standard way. It includes misspelling, typing errors, a forgotten or duplicated character, TE with missing white spaces, unusual acronyms and abbreviations, and all sorts of these problems found in unstructured medical data. An example of this problem in figure 2 is the tag `<TIMEX3 value="PXY" type="DURATION" tid="t2">anos</TIMEX3>`. The word “anos” (years) was split from the number “62”, because there is a missing white space between the comma “,” and the number “62”. A misspelling problem also occurs in tag `<TIMEX3 value="P4Y" type="DURATION" tid="t1">4 anos</TIMEX3>`, where TE “4 anos” (4 years) should be “há 4 anos”, instead of “à 4 anos”.

```
<?xml version="1.0"?>
<!DOCTYPE TimeML SYSTEM "TimeML.dtd">
<TimeML>
<!--<TIMEX3 count="2"/>-->
  pcte xxxxx xx xxxxxx xxxxxxxx xx xxxxxx ,61
  <TIMEX3 value="PXY" type="DURATION" tid="t2">anos</TIMEX3>
  . #s: retorna com exames de ultrassonografia de parte moles com
  laudo de nódulo subcutaneo de provável origem lipomatosa. #e:
  presença de nódulo na fossa supraespinhal. exame laboratorial
  sem alteração. #a: presença de nódulo de origem lipomatosa
  confirmado pelo exame realizado. #c: faço aih relata: tratamento
  de osteoporose e fibromialgia descompensada à
  <TIMEX3 value="P4Y" type="DURATION" tid="t1">4 anos</TIMEX3>
  com suplementação de cálcio. cirurgia para retirada de cisto
  dentário, após implante alérgia na cirurgia, não soube informar
  ao medicamento. labirintite histórico de dois sobrinho e irmão
  com passado de cirurgia para retirada de nódulo lipomatosa.
</TimeML>
```

Figure 2 – Temporal tagging over a noisy clinical report written in Portuguese

While news is the most studied domain, English is the most studied language in the temporal information research area (CAMPOS et al., 2015). However, other languages have also been explored. Languages like Basque (ALTUNA; ARANZABE; ILARRAZA, 2017), French (NZALI; TANNIER; NÉVÉOL, 2015), Chinese (LIU et al., 2016), Portuguese (CRAVEIRO; MACEDO; MADEIRA, 2012) (HAGEGE; BAPTISTA; MAMEDE, 2009), Uyghur (ZHOU; WANG,) and Polish (KOCOŃ; MARCIŃCZUK, 2017) have provided input to temporal information research. In this sense, the scope of this dissertation is to study temporal tagging in noisy clinical reports written in Brazilian Portuguese.

Regardless of domain or language, temporal tagging has been done by three different strategies, the rule-based approach, the machine-learning approach and the hybrid approach (a mix of the other approaches). The most used and successful one has been the rule-based approach. A detailed explanation about these different approaches is provided in chapter 3.

1.1 Motivation

The motivation for this work is the lack of methods of Information Retrieval (IR) to deal with time in real world clinical texts written in Brazilian Portuguese. Clinical texts are usually not well-written, meaning they do not follow the norms of the language. It means that, in many cases, they do not respect the syntax and semantic rules. They also are not logically well-structured, lacking of a well defined start and end. These type of texts usually are not clear to understand. This text style is usually motivated by the health professionals environment stress. This situation was the motivation to create a method to make temporal tagging in noisy clinical reports written in Brazilian Portuguese. Usually works in the literature refer to temporal tagging being performed on texts specifically designed for NLP competitions, these texts being generally well-written and mostly in English.

The study here pursued drew on texts written in Brazilian Portuguese. This research is not the first work to study temporal information written in this specific language. Research on the temporal dimension in Brazilian Portuguese has focused, for example, on the use of finite-state lexical transducers to identify and tag temporal adverbs in Portuguese, specially related to the noun *ano* (year) (BAPTISTA, 2003a). This same strategy was used to handle families of compound temporal adverbs clustering around the noun in Portuguese *ano* (year) (BAPTISTA, 2003b). Craveiro, Macedo and Madeira (2008) proposed a system called PorTexto to recognize TEs in Portuguese, though not in a specific domain, like the clinical one. Different from other systems, it tries to identify TEs in phrases, considering one sentence at a time. The system uses regular expressions to extract temporal expressions. PorTexTo does not normalize the extracted TEs.

Craveiro, Macedo and Madeira (2010) also published an article describing a tool aimed at performing temporal segmentation of documents. The tool performs temporal tagging by using semantically classified temporal patterns (regular expressions) retrieved from a training corpus. Patterns are used in the process of recognition and classification of Portuguese TEs, followed by their annotation in the text. Finally, the normalization step is carried out relying on a set of rules, used to interpret previously annotated TEs and normalize them.

Craveiro, Macedo and Madeira (2012) also proposed a way to enrich an informa-

tion retrieval system with temporal information. They built a testbed system to extract temporal information from Portuguese texts. Their proposal was to achieve temporal characterization of a text collection. They performed TER and normalization. TER is carried out by semantically classified temporal patterns, based on regular expressions. Normalization is performed by a rule-based strategy. The data used by [Craveiro, Macedo and Madeira \(2012\)](#) and by the previously mentioned authors is a subset of the data described in the book by [Mota and Santos \(2008\)](#), which examines an NLP challenge focusing on Portuguese called HAREM. Those data pertain to different domains, however, there being no mention of noisy clinical data.

The work by [Hagege, Baptista and Mamede \(2009\)](#) follows a similar track used in the previous works. They also used the same data and a rule-based strategy. Unlike them, that study aimed at a proper characterization of temporal expressions. It introduced a set of guidelines for TE annotation for Portuguese text (similar to TimeML). In the work by [Ferreira, Teixeira and Cunha \(2010\)](#), TER is one component of the Named Entity Recognition (NER) proposed system. The data input used in their system was from the clinical domain. However, their study did not cover the normalization step.

The work by [Tissot et al. \(2019\)](#) aimed to analyse and normalise imprecise temporal expressions in English and Brazilian Portuguese. Examples of this kind of TEs in Portuguese are: "menos de dois dias" (less than two days) and "algumas semanas" (some weeks). To do so, the authors represented temporal imprecision as membership functions (trapezoidal and hexagonal fuzzy membership functions), which are inspired by human interpretation of TEs. Their work is a new proposal to deal with the normalization of imprecise TEs, which was tested on many corpora, including a clinical one in Brazilian Portuguese. They used statistical regression and machine learning approaches to measure the most suitable model for each type of temporal imprecision rated. Normalisation models they used were able to capture vagueness of imprecise timexes. They did not cope with noisy TEs.

The research by [Li and Patrick \(2012\)](#) mentions noisy clinical texts. Noisy occurrences are corrected by an automated misspelling correction process, which is a general-purpose corrector, not specialized for the clinical domain. The main focus of correction are events and not temporal expressions. The article brings a list of false positive TEs, which bears some similarities with our work. Temporal expressions are extracted by a rule-based approach. The article does not use the TIMEX3 TimeML tag.

As already mentioned, temporal tagging can be done by three distinct approaches: a rule-based approach, a machine learning approach and a hybrid approach. According to [Kreimeyer et al. \(2017\)](#), the majority of NLP systems for capturing and standardizing unstructured clinical information follow the rule-based approach. Based on this and more specifically on results of different approaches used in the clinical domain, this dissertation

opted for a rule-based approach to perform temporal tagging task in noisy clinical text written in Brazilian Portuguese (more details accounting for this decision are presented in chapters 2 and 3). The process of recognizing, extracting and normalizing TEs is called temporal tagging; likewise, the software that perform these tasks is intuitively called temporal tagger. They can be built following different designs, including a rule-based one. In order to fulfill the requirements posed by the proposed method of this dissertation, a temporal tagger must be able to deal with different domains, as the clinical domain shows features shared by different types of text such as news reports, narratives, short text (SMS) and scientific. The tagger must also have an architecture designed to work with different languages other than English, in the case of this study, Brazilian Portuguese. A temporal tagger that met these requirements was HeidelbergTime (STRÖTGEN; GERTZ, 2010), (STRÖTGEN; GERTZ, 2015), which is also the state-of-the-art tool in many articles found in the literature. More details about this tool are provided in chapter 2.

This dissertation builds on data retrieved from a corpus compiled with clinical narratives found in records of Brazilian hospitals. Figure 1 and figure 2 show examples of these narratives. Clinical narratives in the corpus were originally untagged; one of the tasks of this dissertation being tagging them. The training corpus was tagged and those tags were used to create rules in HeidelbergTime. The test corpus was used to test the final version of the proposed method. The corpus was compiled from narratives in records of a Brazilian hospital located in one of the federation state, together with narratives in records of hospitals in another state of the federation. These states are around 1,000 kilometers apart, distance that is taken here as an evidence of distinct contexts of language use by people with distinct cultural, mindset, education level and different ways of thinking and writing.

Regarding works that deal specifically with temporal information in noisy clinical reports written in Brazilian Portuguese, the only work found was the one developed by Tissot and colleagues (TISSOT, 2016). This is focused on the normalization of imprecise TEs. However, as a secondary objective, the work also deals with misspelled (noisy) TEs. The method developed by the authors corrects misspelled TEs by using a dictionary with correctly spelled words which are used in the correction of the misspelled TEs relying on string similarity and phonetic similarity algorithms (TISSOT; PESCHL; FABRO, 2014). Our study shares the same goal, namely to recognize and normalize TEs in noisy clinical reports written in Brazilian Portuguese. Unlike the authors', our study uses a different approach to cope with TEs. Our approach is a pure rule-based one and is designed to work with acronyms and abbreviations as well. This is meant to overcome the limitation (TISSOT; PESCHL; FABRO, 2014; TISSOT, 2016), who do not perform well with acronyms and abbreviations. This limitation is related to the string similarity and phonetic similarity algorithms used in their method, which is not successful in dealing with short TEs. We had originally intended to compare their results with ours. Yet, neither

the dataset nor the source code used by them were made available.

This dissertation brings an innovative approach. It is the first work dealing with temporal tagging of noisy clinical texts written in Brazilian Portuguese using exclusively a rule-based approach. Based on [Ferro et al. \(2001\)](#), who argues that “If a human can determine a value for the temporal expression, it should be tagged”, a further contribution of this dissertation is the recognition of specific TEs that construe time in the clinical field in particular, such as “1x/dia” (once a day), “12/12 h” (each 12 hours), “manhã, tarde e a noite” (morning, afternoon and evening - three times a day), “cp no almoço” (a tablet at lunchtime), “mg/dia” (once a day), “tid” (three times a day) among others that were catalogued. Naturally, noisy versions of the TEs were also annotated. This is certainly an advantage of a rule-based approach.

1.2 Objectives

The main objective of this dissertation is to develop a method to extract and normalize temporal expressions in noisy medical records written in Brazilian Portuguese using a rule-based approach.

Specific objectives:

- Annotate and normalize TEs from a noisy clinical corpus written in Brazilian Portuguese;
- Transform TEs in Portuguese into rules to a rule-based temporal tagger;
- Evaluate the results of the temporal tagger with the new rules.

1.3 Hypothesis of the work

The hypotheses examined in this dissertation are:

- H1) Noisy TEs from clinical text improve the assertiveness of a temporal tagger when the TEs are transformed into rules of this tagger;
- H2) The creation of new rules based on parts (n-gram) of noisy TEs increase the quality of a temporal tagger;

Hypotheses were measured by three metrics widely used in the temporal information research area: precision, recall (sensitivity) and f-measure (also called f-score and f1 score) ([MANI et al., 2006](#)), ([PUSTEJOVSKY; VERHAGEN, 2009](#)), ([VERHAGEN et al., 2009](#)), ([REEVES et al., 2013](#)).

1.4 Scientific Contribution

The main contribution of this research is the proposal of a method to extract and normalize TEs in noisy medical records written in Brazilian Portuguese, using a rule-based temporal tagger.

A second contribution of this work is the annotation of TEs in a real world clinical dataset in Brazilian Portuguese. The annotation of the corpus was made following scientific standards by professionals of the healthcare area.

1.5 Scope

The scope of this research is limited to extraction and normalization of TEs in noisy clinical texts typed in Brazilian Portuguese through a rule-based temporal tagger.

1.6 Document Organization

This document is organized in 9 chapters in addition to this one. Chapter 2 presents the basic concepts which are necessary to comprehend this work. Chapter 3 presents the state-of-the-art of this study. Chapter 4 presents the methodological procedures needed in this work. Chapter 5 brings light to the proposed method in this document. Chapter 6 presents the dataset annotation which was necessary to accomplish this dissertation. Chapter 7 details experiments and results of our study. Chapter 8 discusses results and chapter 9 presents the conclusion of our study.

2 Basic concepts

The objective of this chapter is to present a general review about temporal information extraction, the concepts related to it and the tagger used in this study. This research area is concerned about detecting and interpreting temporal expressions, events and temporal relations between events and between temporal expressions and events (VERHAGEN et al., 2009). Temporal tagging (extraction and normalization of temporal expressions) is the focus of this dissertation; its details are discussed in section 2.2. However, as noisy TEs are part of this research, it is important to present formally what the word "noisy" means in this study.

2.1 The word "noisy"

In this study, the word noisy indicates any type of text which is a temporal expression (TE) produced in a non-regular pattern of language or non-standard way. It includes misspelling, typing errors, a forgotten or duplicated character, TE with missing white spaces, unusual acronyms and abbreviations, and all sorts of these problems found in unstructured medical data.

2.2 Temporal Tagging

Temporal tagging can be considered as a particular type of Named Entity Recognition (NER). Person, organization and location are the standard named entity types. However, numeric expressions including time, date, money and percent expressions are also named entity types (NADEAU; SEKINE, 2007). Depending on the need and context, genes and proteins, numbers and temporal expressions can also be clustered as NER (STRÖTGEN; GERTZ, 2016). From a general point of view, temporal tagging can be divided in two steps, extraction and normalization.

Both steps are performed on a temporal expression. This is an expression found in a text which is a date or a time at any level of granularity like ("15-12-2013", "yesterday", "the last 18 months"). It can also be an expression referring to a duration (e.g., "five years", "three weeks") or an expression referring to the periodical aspect of an event (e.g. "every Wednesday", "once a year"). The extraction step is the moment when the temporal expressions in the text are recognized and extracted. In the normalization step, the temporal expression is changed to a standard format. After the normalization step, the first two examples are changed to "2013-12-15" and "2013-12-14". This happens because the temporal expression "yesterday" is relative (related) to 15-12-2013 and both were

normalized (to the same pattern). When a temporal expression is already normalized, it can be called “the value of a temporal expression” (STRÖTGEN; GERTZ, 2016). Figure 3 shows examples of extraction and normalization of temporal expressions.

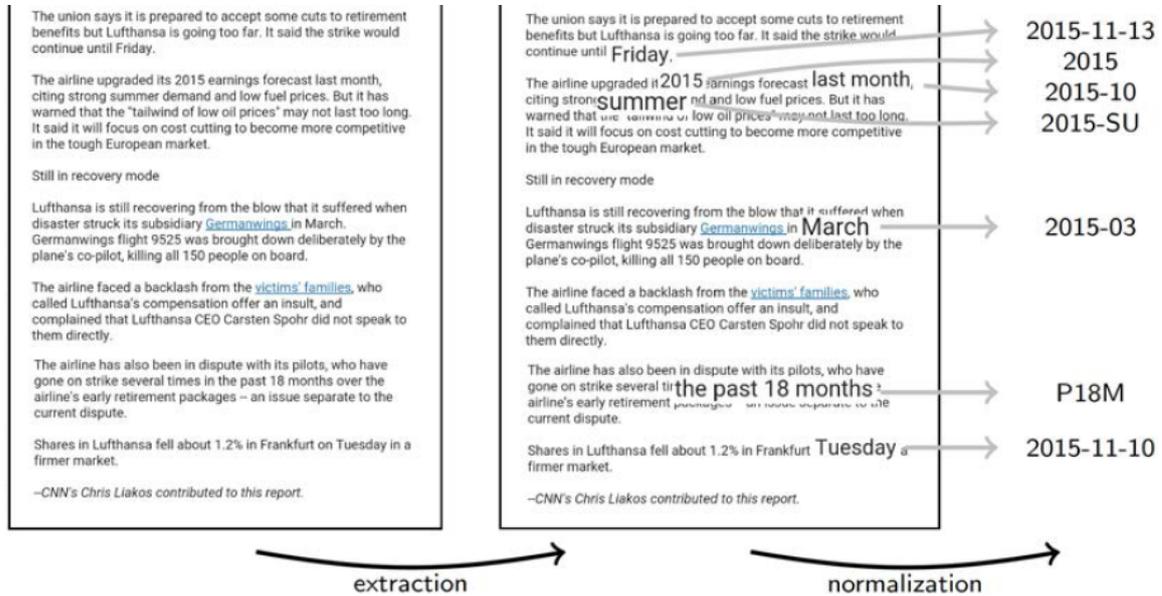


Figure 3 – Extraction and normalization of temporal expressions (Extracted from Strötgen and Gertz (2016)).

A temporal information can still be singled out by three key characteristics: “temporal information is well defined”, “temporal information can be normalized” and “temporal information can be organized hierarchically” (ALONSO et al., 2011).

A temporal information is well defined when, given two points in time or two intervals, the relationship between them can be recognized. This relationship can be of the types: before, overlap or after. Figure 4 shows some of these relationships. The whole set of relationships was proposed by Allen (1983).

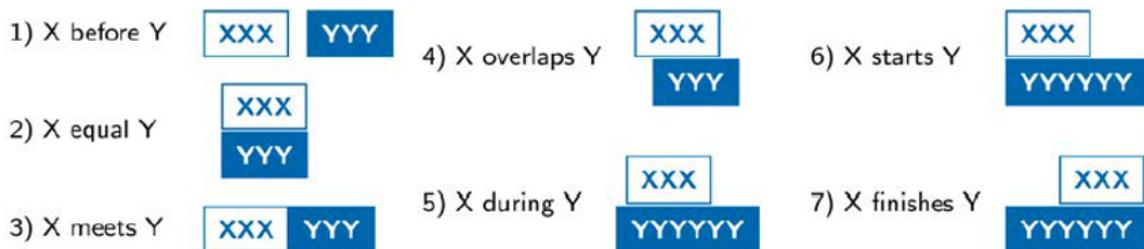


Figure 4 – The main types of well-defined relationships between two points in time (Extracted from Strötgen and Gertz (2016)).

In addition to being well defined, temporal information must be normalized. This feature was already mentioned above. Regardless of terms or language was used, for temporal tagging to be effective, it is critical that temporal information be standardized.

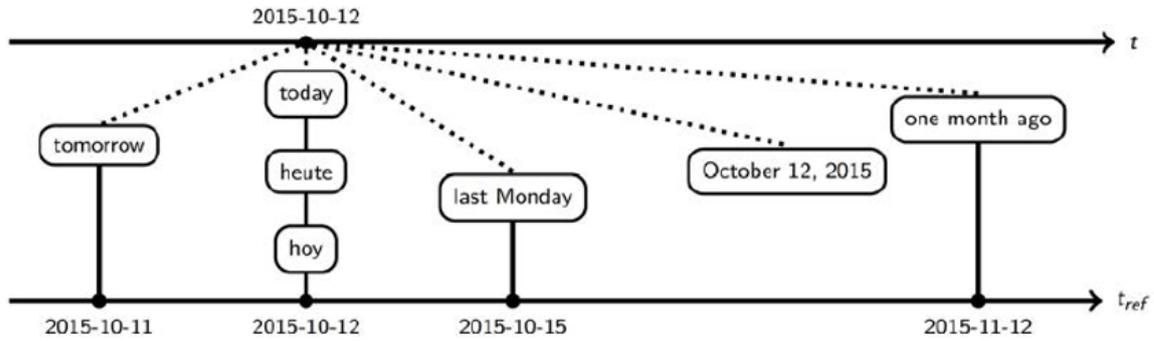


Figure 5 – Temporal information normalization process (Extracted from [Strötgen and Gertz \(2016\)](#)).

Figure 5 shows the normalized date 2015-10-12 as a reference for all other normalized temporal expressions. The normalization must be done independent of language, it is what happens with the German term “heute” and the Spanish term “hoy”, that are equivalent to “today” in English. In addition to being standardized, temporal information also needs to be ordered hierarchically.

The hierarchical ordering of a temporal information deals with the granularity of a temporal expression. The granularity may be from the day point of view, for example “November 2, 2017”. It can be from the month point of view, as in “November 2017”. It can also be a granularity from the year point of view, as in “2017”. There are several other ways for treating temporal expressions under the granularity bias, such as decade, minute, hour, nanosecond etc. The hierarchy depends on the interest of using temporal information in a given context. What matters is that it can be ordered hierarchically. It is important to realize that the month granularity, for example, includes various day granularities. Just as an hour granularity includes many minutes and seconds in itself. A hierarchical ordering of temporal expressions can be useful, for example, to define a form of normalization of a text. Figure 6 illustrates the hierarchical ordering of temporal expressions.

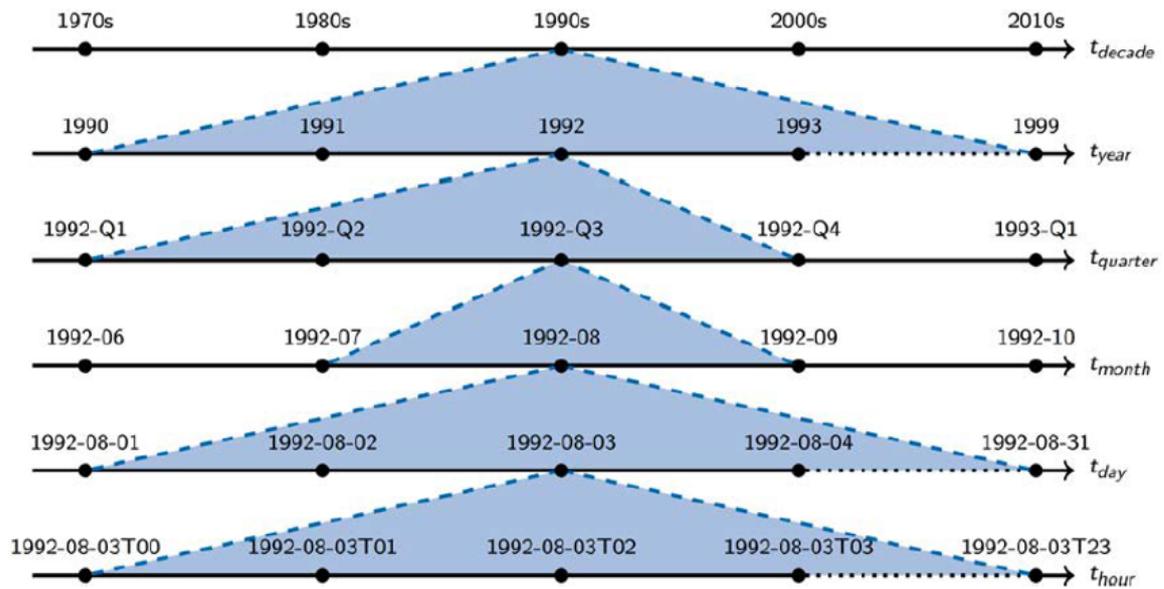


Figure 6 – Temporal information ordered hierarchically (Extracted from Strötgen and Gertz (2016)).

To ease the normalization process, it is possible to consider a strategy to convert all entities of date and time to the smallest unit of time found in the document being tagged, but it would erase the semantics of expressions. For example, it would be strange to say that a treatment has lasted for seconds or minutes (MADKOUR; BENHADDOU; TAO, 2016).

In addition to the three key features presented, there are different types of temporal expressions, namely date expressions, time expressions, duration expressions and set expressions. These expressions have been defined drawing on temporal tagging languages, mostly on one of them, TimeML (PUSTEJOVSKY et al., 2003), (PUSTEJOVSKY et al., 2005), (MADKOUR; BENHADDOU; TAO, 2016).

- Date expression: this type of expression refers to a point in time of the granularity day, like in “November 2, 2017”, “November 2017” or “2017”.
- Time expression: refers to a point in time of any granularity smaller than day, for example, parts of the day. “Thursday morning” and “3:30 p.m.” are examples of time expressions.
- Duration expressions: this type of expression deals with the length of an interval. These intervals can be of different granularities, for example, “5 hours”, “5 days”, “3 years”, etc. A duration expression can also specify the point in time where an interval starts and ends.
- Set expressions: this kind of expression refers to the periodic aspect of an event. It serves to describe a set of times, dates, or frequency within a time interval that an

event occurs. Examples of set expression are: “every Thursday” and “three times a week”.

In addition to the types of temporal expressions, there are other important concepts within the area of temporal tagging. Among these concepts the realization of a temporal expression is important. In this work we adopt four realizations of temporal expressions defined by [Strötgen \(2015\)](#). These realizations are: explicit expressions, implicit expressions, relative expressions and underspecified expressions. The realization type of a temporal expression is directly related to its normalization complexity.

- Explicit expressions: when a temporal expression belongs to this type of realization, it has all information to be normalized. These expressions may be of the date or time type, they can treat a granularity equal to or greater than one day, or a granularity less than one day, respectively. For example, temporal expressions of day granularity as in, “November 2, 2017” or “November 2017” can be normalized to 2017-11-02 and 2017-11, respectively.
- Implicit expressions: the normalization of implicit expressions is facilitated by knowing the date of the implicit expression. This category is designed specifically for named dates, an example of this type of temporal expression is Christmas, which is always celebrated on December 25. Thus, when the 2017 Christmas term is found in a text, for example, it can be normalized to 2017-12-25. However, there are implicit expressions that require a calculation that precedes its normalization. This is the case of Mother’s Day in Brazil, which is always celebrated on the second Sunday of May. Therefore, the specific date of the holiday is not fixed. In this way, when an NLP system finds in a text the term "Mother’s Day", it calculates the date of the second Sunday of May in the respective year. In the case of 2017, the normalization of Mother’s Day would be 2017-05-14.
- Relative expressions: As can be seen in the two previous types of temporal expressions, they do not need a context to be normalized. This does not happen with relative and unspecified temporal expressions. When terms like “today”, “next month” and “five days later” are found, they need a reference time to be normalized. For example, in the text “He was discharged two days ago, but today he returned to the hospital”, first is necessary to know when is “today”, and then, the term “two days ago” can be calculated in relation to “today”. Suppose the NLP system is able to identify that “today” is “October 3, 2016”, in this case, “two days ago” must be “October 1, 2016”. Once the temporal expressions are identified, its normalization would be 2016-10-03 and 2016-10-01 respectively. These are examples of relative expressions.

- **Underspecified expressions:** in this type of temporal expression, the context for normalization is also important. However, this type of temporal expression groups expressions that have missing data. In these cases, it is not possible to find a reference time in the context. Examples of this type of expression are the terms “November” and “November 15”. Assuming that the context where these terms are inserted does not present the year information, the normalization of these expressions would be XXXX-11 and XXXX-11-15 respectively.

Figure 7 shows an example of the four realization types.

Lufthansa tries to force striking staff back to work **November 10, 2015**

The carrier confirmed **Tuesday** it had applied for a German court order to end the strike, which has now caused roughly 1,900 canceled flights since **Friday**.

...
 “There was no offer to reject **yesterday** ...”.

...
 The airline upgraded its **2015** earnings forecast **last month** ...

...
 Lufthansa is still recovering from the blow that it suffered when disaster struck its subsidiary Germanwings in **March**

<http://money.cnn.com/2015/11/10/news/companies/lufthansa-strike-court-injunction> [last accessed: Nov 21, 2015]

Two days ago , Lufthansa also announced that there will be no excuse for a strike on **Thanksgiving 2015** .

explicit temporal expression

underspecified temporal expression

relative temporal expression

implicit temporal expression

Figure 7 – Four realization types of temporal expressions (Extracted from Strötgen and Gertz (2016)).

In addition to the concepts discussed above, another detail within the universe of temporal information is important, it is the annotation pattern. When a temporal expression is identified, extracted and normalized, it must be annotated or marked in a pattern that allows its use as input to another system. There are two main patterns of annotation of temporal information in the literature, they are the TIDES TIMEX2 (FERRO et al., 2001) and the TimeML (PUSTEJOVSKY et al., 2003), this second uses TIMEX3 tags for temporal expressions annotation. In both cases, the normalization is done according to ISO 8601. Each one of the patterns will be presented in the next paragraphs.

According to Strötgen and Gertz (2016), TIDEX TIMEX2 was the first annotation guide defined with enough detail to be accepted as a standard. TIDEX TIMEX2 has two important principles for an annotation standard from the point of view of standardization and understanding by humans. The first principle reads “if a human can determine the value for a temporal expression, that expression should be marked.” The second principle

posits that an annotation “must be based on internal evidence of the document being annotated” [Ferro et al. \(2001\)](#). The attributes of TIMEX2 are presented in the table 1.

Table 1 – TIMEX2 attributes (Extracted from [Ferro et al. \(2005\)](#))

Attribute	Function	Example
VAL	Contains a normalized form of the date/time	VAL=“1964-10-16”
MOD	Captures temporal modifiers	MOD=“APPROX”
ANCHOR_VAL	Contains a normalized form of an anchoring date/time	ANCHOR_VAL=“1964-10-16”
ANCHOR_DIR	Captures the relative direction/orientation between VAL and ANCHOR_VAL	ANCHOR_DIR=“BEFORE”
SET	Identifies expressions denoting sets of times	SET=“YES”
COMMENT	Contains any comments the annotator wants to add	COMMENT=“context garbled”

As previously said, TimeML became an ISO, also called ISO-TimeML (ISO 24617-1:2012). TimeML is based on the TIDES standard ([PUSTEJOVSKY et al., 2003](#)), it is considered TIDES natural evolution. TimeML is the current temporal tagging annotation standard. Unlike TIDES standard, which has tags only for temporal expressions, TimeML has broader and more general temporal information coverage. This standard contains tags for event annotation, annotation of temporal links (temporal relations), etc., as well as TIMEX3 tags for temporal expressions ([PUSTEJOVSKY et al., 2003](#)). The attributes of TIMEX3 tag are ([GROUP et al., 2009](#)), ([SAURI et al., 2006](#)).

1. Timex ID number (tid) - Mandatory attribute. It is an unique idenfier generated automatically by the annotation tool.
2. type - non-optional attribute. This attribute is used to define if a temporal expression is a date, time, duration or set
3. value - it corresponds to the normalized form of a temporal expression.
4. mod - used to capture temporal modifiers.
5. quant and freq - these modifiers specify the quantity and frequency of set expressions.
6. beginpoint and endpoint - they are used to anchor the beginning and the end of a duration.

Some examples of the use of the above-mentioned attributes for times of day and dates are shown below ([GROUP et al., 2009](#)).

```
<TIMEX3 tid="t1" type="TIME" value="T24:00">
twelve o'clock midnight
</TIMEX3>
```

```
<TIMEX3 tid="t4" type="TIME" value="1984-01-03T12:00">
twelve o'clock January 3, 1984
</TIMEX3>
```

```
<TIMEX3 tid="t5" type="DATE" value="XXXX-12-02">
the second of December
</TIMEX3>
```

```
<TIMEX3 tid="t6" type="DATE" value="1964-SU">
summer of 1964
</TIMEX3>
```

The examples that follow refer to durations.

```
<TIMEX3 tid="t1" type="DURATION" value="P4M">
4 months
</TIMEX3>
```

for

```
<TIMEX3 tid="t1" type="DURATION" value="PT20M">
20 minutes
</TIMEX3>
```

```
<TIMEX3 tid="t2" type="DATE" value="XXXX-07-08">
last Monday
</TIMEX3>
```

We'll be on a break from

```
<TIMEX3 tid="t1" type="DATE" value="XXXX-07-01">
July, 1
</TIMEX3>
```

to

```
<TIMEX3 tid="t2" type="DATE" value="XXXX-08-15">
August, 15
</TIMEX3>
```

```
<TIMEX3 tid="t3" type="DURATION" value="P46D"/>
```

Examples of use of the `mod` attribute are presented in the table 2.

Table 2 – Values for MOD attribute (Extracted from [Group et al. \(2009\)](#))

Durations	LESS_THAN MORE_THAN EQUAL_OR_LESS EQUAL_OR_MORE	less than 2 hours long more than 5 minutes no more than 10 days at least 10 days
-----------	----------------------------------------------------------	-------------------------------------------------------------------------------------------

Examples of use of `quant` and `freq` attributes in set expressions can be seen below ([GROUP et al., 2009](#)).

```
<TIMEX3 tid="t1" type="SET" value="P1W" freq="2X">
twice a week
</TIMEX3>
```

```
<TIMEX3 tid="t1" type="SET" value="P1W" quant="EACH"
freq="3d">
3 days each week
</TIMEX3>
```

```
<TIMEX3 tid="t1" type="SET" value="XXXX-10"
quant="EVERY">
every October
</TIMEX3>
```

The next paragraph presents an example of use of attributes `beginpoint` and `endpoint` ([GROUP et al., 2009](#)).

```
He left between
<TIMEX3 tid="t1" type="TIME" value="T18:00">
6:00 p.m.
</TIMEX3>
and
<TIMEX3 tid="t2" type="TIME" value="T20:00">
8:00 p.m.
</TIMEX3>
```

```
<TIMEX3 tid="t3" type="DURATION" value="P2H"  
beginPoint="t1" endPoint="t2"/>
```

2.3 Metrics

Dealing with temporal information is a complex scenario, which involves many features and variables. An important feature of temporal tagging is how the process is evaluated. In other words, how it is possible, from a metric point of view, to say that the temporal tagging was well done. When temporal taggers (a software that performs temporal tagging) is evaluated, it is done intrinsically, using manually annotated data and evaluating a temporal tagger's extraction and normalization based on manually annotated corpora [Strötgen and Gertz \(2016\)](#).

A confusion matrix can be used to describe how a system performing temporal tagging works well when compared to a gold standard. The results of a confusion matrix will always fall into one of the four cases ([MANNING; SCHÜTZE et al., 1999](#)).

- **True positive (TP)**: a temporal expression annotated by the system and in the gold standard;
- **True negative (TN)**: a temporal expression neither annotated by the system nor in the gold standard;
- **False positive (FP)**: a temporal expression annotated by the system, but not present in the gold standard;
- **False negative (FN)**: a temporal expression not annotated by the system, but present in the gold standard;

It is important to be aware that, as many temporal expressions are made of more than one word, it is also interesting to distinguish between strict and relaxed matching. In a strict matching approach, when the temporal expression “Wednesday afternoon” exists in the gold standard, the temporal tagger only will recognize this temporal expression if it is written exactly the way it is in the gold standard. However, in a relaxed approach, if the system finds the word “Wednesday” in the analyzed text, it will be considered a match with the gold standard. According to [Strötgen and Gertz \(2016\)](#), the relaxed matching with correct normalization is usually considered as the most important evaluation measure. Once strict and relaxed issue was explained, an example of the confusion matrix is shown in table 3.

Table 3 – The setup of a confusion matrix, used to evaluate temporal taggers (Extracted from [Strötgen and Gertz \(2016\)](#))

System Prediction	Gold Standard (Ground Truth)	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

The data necessary to build a confusion matrix are directly or indirectly the commodities for all metrics which measure different aspects of the temporal tagging task. The extraction or normalization steps of the temporal tagging task can have its efficiency measured by a confusion matrix. Considering only the extraction phase, true positive instances are the ones correctly extracted by a temporal tagger. On the other hand, for normalization, true positives are only the instances which were correctly extracted and normalized. Usually, precision, recall and f1-score are the measures to evaluate a system that performs a temporal tagging task ([STRÖTGEN; GERTZ, 2016](#)), ([MANNING; SCHÜTZE et al., 1999](#)).

The precision measure is used to indicate how many expressions extracted by the tagger are correct. When all instances tagged as positive by the system are in fact correct, the precision result will be 1. On the contrary, if all predictions made by the tagger as correct instances are, in fact, incorrect, the precision result will be 0. Figure 8 shows the precision formula.

$$precision = \frac{TP}{TP + FP}$$

Figure 8 – Precision formula

The recall measure is the precision measure counterpart. It indicates how many expressions that should be extracted are correctly extracted by the tagger. Recall equal to 1 means that all instances which should be tagged as positive have indeed been tagged as positive by the tagger. Recall equal to 0 indicates that all instances that should not be marked as positive have been marked as positive by the system. Figure 9 presents the recall formula.

$$recall = \frac{TP}{TP + FN}$$

Figure 9 – Recall formula

Precision and recall considered alone are biased measures. An example makes it clear. Marking all instances (of a text) as positive results in a recall of 1, while marking only a single instance correctly as positive results in a precision of 1. The example shows

that there is a need of balance between precision and recall. To reach this balance, the f-measure (also called f-score, f1-score or f1) was developed, it calculates the balanced harmonic mean. Figure 10 shows its formula.

$$f_1\text{-score} = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$

Figure 10 – F1-measure formula

From a practical perspective, [Strötgen and Gertz \(2016\)](#) highlight that there is a tendency that the focus of the evaluation is put on the TIMEX3 attributes “type” and “value”. Between these two attributes, the “value” is the most important for evaluation purposes.

In order to evaluate the normalization quality of a temporal tagger, the accuracy measure is an alternative. Precision and accuracy measures can seem to be equal, but are not. The precision measure considers in its calculation only TP and TN. On the other hand, accuracy calculates all decisions (TP, TN, FP and FN), whether they are marked as positive or negative in the gold standard. Accuracy formula is presented in figure 11.

$$\textit{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Figure 11 – Accuracy formula

Accuracy is not appropriate to measure the temporal expression extraction, because the majority of instances are negative (not temporal expressions). In contrast, accuracy can be used to evaluate the normalization task. However, there is a better measure for this situation, it is the normalization-accuracy measure. This measure only takes into account the temporal expressions correctly extracted by a temporal tagger. Normalization-accuracy evaluates if the temporal expressions extracted correctly by the tagger were also normalized in a proper way. Its formula is presented in figure 12.

$$\textit{normalization-accuracy} = \frac{\textit{correctly normalized expressions}}{\textit{correctly extracted expressions}}$$

Figure 12 – Normalization-accuracy formula

When two temporal taggers must be compared, its extraction and normalization quality have to be considered. In this case, the normalization-accuracy measure can be useful ([STRÖTGEN; GERTZ, 2016](#)).

2.4 HeidelTime

Among all rule-based temporal taggers analyzed in this study, HeidelTime was the only one that met the requirements to work with text from the clinical domain and with the Portuguese language. It has also been referred to as the state-of-the-art temporal tagger by many articles in the literature.

HeidelTime¹ is a multilingual, domain-sensitive temporal tagger developed at the Database Systems Research Group at Heidelberg University in Germany. It was designed to extract TEs from documents and normalize them according to the TIMEX3 annotation standard. HeidelTime contains manually created resources for 13 languages and automatically created resources for more than 200 languages.

The tagger was created to fulfill a lack of publicly available temporal taggers, especially to be used for processing different languages and domains in extraction and normalization of TEs. The details of HeidelTime are presented according to [Strötgen and Gertz \(2013\)](#). The authors stress that, even if there are promising machine learning approaches for the extraction of temporal expressions (TEs), HeidelTime was developed as a rule-based system for the following reasons:

1. The divergence of TEs is limited compared to other named entity recognition and normalization tasks, e.g., the number of persons and organizations as well as the variety of names referring to these entities are probably infinite;
2. The normalization task is hardly solvable without using rules;
3. Resources for additional languages can be added without the need of an annotated corpus;

In the next sub-sections, HeidelTime will be looked at closer.

2.4.1 HeidelTime's architecture

An overview of HeidelTime's architecture is presented in figure 13. The system was developed with a strict separation between the algorithmic part (code, resources for patterns, rules) and the normalization information. It was organized in a modular way. This strategy allows HeidelTime to be extensible. From a multilingual perspective, only its resources are language-dependent, what makes the integration of a new language easier.

¹ <https://github.com/HeidelTime/heideltime>

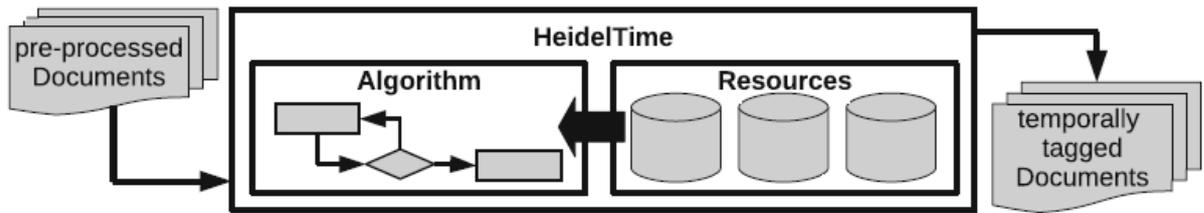


Figure 13 – HeidelTime architecture with algorithm (source code) and resources. Extracted from [Strötgen and Gertz \(2013\)](#).

In the extraction task, HeidelTime mainly uses regular expressions that can use pattern resources. Nevertheless, other constraints can be used as well (part-of-speech tag of a specific token in the expression itself or before or after the TE). In the normalization task, normalization resources containing mappings between an expression and its value in a standard format are used. The next sub-section explains HeidelTime’s resources.

2.4.2 HeidelTime’s resources

HeidelTime’s resources are organized in a tree-structure ([STRÖTGEN; GERTZ, 2013](#)). In the first level there are folders for each language present in the system. Within each language folder, there are at least three folders in a modular way, they are: rules, patterns and normalization. Within these directories, every resource item is represented as a file in which one can easily modify the resource or include comments and examples without influencing the resource itself. Figure 14 illustrates the resources.

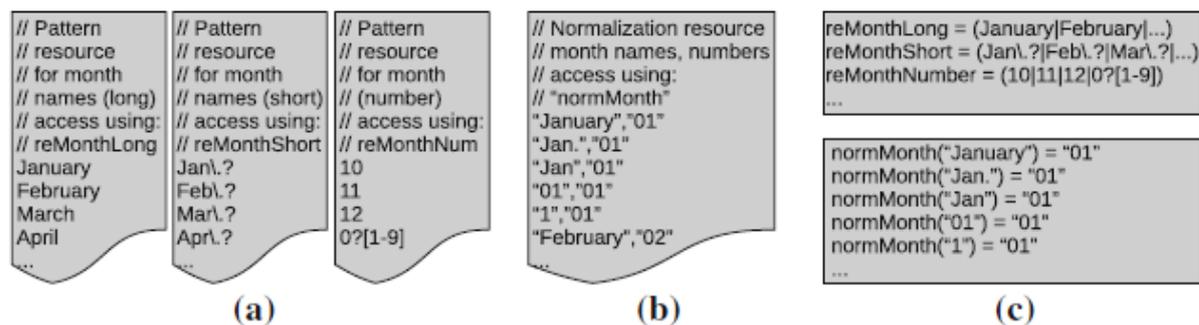


Figure 14 – Pattern resource files for different expressions for month names and numbers (a) and normalization resource file for month expressions (b). (c) represents how their information is used in HeidelTime after being read and translated by HeidelTime’s resource interpreter. Extracted from [Strötgen and Gertz \(2013\)](#).

- Pattern resources: They are used to create regular expressions, that can be accessed by every rule. This allows to use category names instead of listing all items every time the category is needed in a rule. For example, there are patterns for month

names, names of weekdays, and number words. The pattern resource files contain one disjunct per line and the regular expression is built by HeidelbergTime's resource interpreter when reading the resources. Figure 14a shows examples of pattern resource files, and figure 14c (upper part) how they are translated by the resource interpreter.

- Normalization resources: These resources contain normalized values of expressions included in the pattern resources. These values often correspond to the TimeML standard for temporal information (called TIMEX3). They are used when HeidelbergTime assigns a value to a TE, i.e., when interpreting the TE. The normalization resource files are read by HeidelbergTime's resource interpreter, and for every file, a hash map is created. The files contain one key/value pair in each line. An example of a normalization resource file is given in figure 14b. The HeidelbergTime normalization flow is showed in the lower part of figure 14c.
- Rule resources: These resources contain the rules for the extraction and the normalization of TEs. There is one file for every TE type (date, time, duration and set). In the extraction part and the normalization part of the rules, one can use the pattern resources and the normalization resources, respectively. In addition, it is possible to define further constraints, such as specific part-of-speech tags at a specific position, or modify the extent of a TE. Similar to other resources, the rule resources are read by HeidelbergTime's resource interpreter and hash maps are created for the extraction, the normalization, and for all further constraints.

The next sub-section presents the syntax of the rule language use in HeidelbergTime.

2.4.3 HeidelbergTime's rule syntax

For HeidelbergTime, Strötgen and Gertz (2013) consider every TE as a three-tuple $te_i = \langle e_i, t_i, s_i \rangle$, with the expression itself (e_i), the type of the expression (t_i , with $t_i \in \{date, time, duration, set\}$), and the normalized semantic attributes of the expression (s_i). Note that s_i does not only consist of the TIMEX3 attribute value, but of all attributes that are subject to normalization, e.g., the *mod* attribute. The goal of HeidelbergTime is, for each TE te_i in a document, to identify the expression e_i and its type t_i and to correctly normalize its semantics s_i . In order to accomplish HeidelbergTime scope, a rule syntax was developed.

- RULENAME: Give a name to each rule allows to retrace which rule extracted each TE. It is useful for several tasks, e.g., calculating statistics of the occurrences of different TEs, and performing a detailed error analysis and improving rules.

- **EXTRACTION**: Every rule contains an extraction part describing the regular expression pattern that has to be matched. This is the part where pattern resources are applied. It is possible to use parentheses to group parts of the expression, which is important for the normalization of the expressions.
- **NORM_VALUE**: This part defines the normalized value of an expression. One can use the normalization resources described in the previous section and refer to parts of an expression using the groups defined in the extraction part of the rule. In addition, the following functions can be applied:
 - **SUBSTRING(x,i,j)**: returns a substring of the string *x* starting at character position *i* and having the length *j*.
 - **LOWERCASE(x)**: converts all characters of the string *x* to lower case.
 - **UPPERCASE(x)**: converts all characters of the string *x* to upper case.
 - **SUM(x,y)**: adds the integer *y* to the integer *x*.

Beyond the three parts of the rules already described (*rulename*, *extraction*, *norm_value*), for some linguistic phenomena, it is necessary to specify further constraints to correctly extract a specific TE. The following parts can be added to a rule ²:

- **POS_CONSTRAINT(group(x):y)**: part-of-speech of group *x* is *y*
- **OFFSET(group(x)-group(y))**: offset begins at beginning of group *x* and ends at the end of group *y*
- **NORM_MOD**: it is where the attribute *mod* is set
- **NORM_QUANT**: it is where the attribute *quant* is set
- **NORM_FREQ**: it is where the attribute *freq* is set

In the next paragraphs examples of Heidelberg's rules using the aforementioned parts and functions are presented.

Simple example rule 1³

```
RULENAME="date_simple1",
EXTRACTION="%reMonthLong %reDayNumber, %reYear4Digit",
NORM_VALUE="group(3)-%normMonth(group(1))-%normDay(group(2))"
```

² <https://github.com/HeidelTime/heideltime/wiki/Developing-Resources>

³ <https://github.com/HeidelTime/heideltime/wiki/Developing-Resources>

Through the % sign is possible to refer to a pattern resource or a normalization resource in the extraction and norm_value parts of a rule, respectively. Assuming that *reMonthLong* contains all month names, *reDayNumber* a regular expression for the numbers between 1 and 31, and *reYear4Digit* a four digit number, then this rule matches expressions such as:

- “January 9, 2011” Since every pattern resource counts as one group, the following groups are available for the normalization process:
 - group(1) = “January”
 - group(2) = “9”
 - group(3) = “2011” As defined by TimeML, this expression has to be normalized to “2011-01-09”. The normalized expression consists of the year as it is mentioned in the expression (*group(3)*) followed by a “-”. The next part is the month. However, the month name January has to be normalized to “01”. Thus, the normalization resource *normMonth* is called to normalize the *group(1)* expression. Finally, the day (9) is normalized to “09” using the *normDay* normalization resource.

The normalized expression is thus “2011-01-09”.

Simple example rule 2

```
RULENAME="date_simple2",
EXTRACTION="(reMonthLong|reMonthShort) reDayNumber, reYear4Digit",
NORM_VALUE="group(5)-normMonth(group(1))-normDay(group(4))"
```

Assuming that *reMonthShort* contains abbreviated month names, this rule additionally matches expressions such as:

- “Mar. 11, 1982” Note that the group information changed, because an additional pattern resource and a new parentheses pair was added. This results in the following group information:
 - group(1) = “Mar.”
 - group(2) = null
 - group(3) = “Mar.”
 - group(4) = “11”

- `group(5)`="1982" Note that depending on whether `reMonthLong` or `reMonthShort` matches the expression either `group(2)` or `group(3)` equals null. Independent on which resource matches, `group(1)` contains the expression. Thus in the `norm_value` part `group(1)` is used with the `%normMonth` resource.

Complex example rule 1⁴

```
RULENAME="date_complex3",
EXTRACTION="%rePartWords([ ]?)%reYear4Digit",
NORM_VALUE="group(3)",
NORM_MOD="%normPartWords(group(1))"
```

Assuming `rePartWords` matches expressions such as “the beginning of”, this rule matches TEs such as “the beginning of 2001”. The following group information is available:

- `group(1)`="the beginning of"
- `group(2)`=" "
- `group(3)`="2001" This rule sets the value to “2001” and additionally sets the modification attribute `MOD` to “START” since the normalization resource `normPartWords` contains a line “the beginning of”, “START”.

Complex example rule 2

```
RULENAME="date_complex_negative",
EXTRACTION="%reYear4Digit ([\S]+)",
NORM_VALUE="REMOVE",
POS_CONSTRAINT="group(2):NNS:"
```

This rule is an example for negative rules. Their task is to prevent phrases to be matched as TEs. Assume a rule matching every four digit number as a TE of the type year. This results in many false matches since not every four digit number refers to a year expression. The task of the negative rule is to block the four digit number to be extracted as a TE if it is unlikely to be a TE. This rule matches expressions starting with a four digit number followed by any kind of token. Thus, expressions such as “2000 shoes” are matched by this rule and the following group information is available:

⁴ <https://github.com/HeidelTime/heideltime/wiki/Developing-Resources>

- `group(1)`=“2000”
- `group(2)`=“shoes” In this rule, an additional constraint is set, namely that `group(2)` has the part-of-speech *NNS* (plural noun). This constraint is satisfied in the given example. In the norm value part of the rule the value is set to “REMOVE”. In Heidelberg’s source code, this “REMOVE” is interpreted in such a way that this phrase and all sub-phrases are blocked as TEs.

Underspecified and relative expressions⁵

For underspecified and implicit expressions, one can set the value in an underspecified way.

- `UNDEF-(this|next|last)-%normUnit(x)`
- `UNDEF-this-%normUnit(x)-REST`
- `UNDEF-REF-%normUnit(x)-REST`
- `UNDEF-REFUNIT-%normUnit(x)-REST`

REST may either be a part of the expression, or additionally contain a calculation function.

Relative example rule

```
RULENAME="date_relative_1",
EXTRACTION="([\d]+) %reUnit ago",
NORM VALUE="UNDEF-this-%normUnit(group(2))-MINUS-group(1)"
```

This rule matches expressions such as “20 years ago” with the following group information being available for the normalization process:

- `group(1)`=“20”
- `group(2)`=“years” The expression is such normalized to “UNDEF-this-year-MINUS-20”. The calculation is performed during Heidelberg’s disambiguation phase.

⁵ <https://github.com/HeidelTime/heideltime/wiki/Developing-Resources>

2.4.4 HeidelTime’s algorithm

In this section the explanation of HeidelTime’s algorithm is done based on the paper of the creators of the system, [Strötgen and Gertz \(2013\)](#). The overall functioning of the algorithm is presented in figure 15.

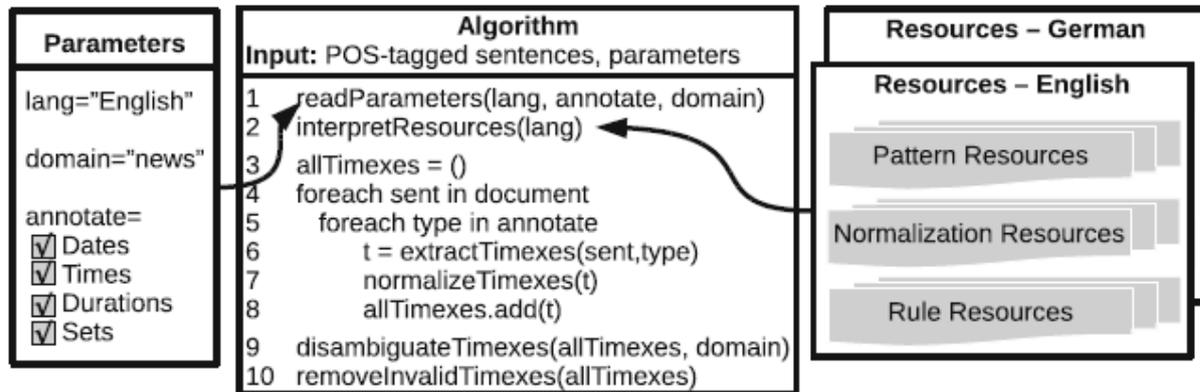


Figure 15 – HeidelTime’s algorithm overview. Extracted from [Strötgen and Gertz \(2013\)](#).

As its first step HeidelTime expects as input part-of-speech tagged sentences and user-specified parameters defining which types of expressions are to be annotated (parameter *annotate*) and which language and domain are used (parameters *lang* and *domain*, respectively). In an initialization phase, the parameters are read (line 1) and the resources of the corresponding language are interpreted by HeidelTime’s resource interpreter (line 2). Then, HeidelTime performs the extraction and normalization of TEs by running the following phases: (1) the extraction phase, (2) the normalization phase, (3) the disambiguation phase, and (4) the cleaning phase. In Fig. 15, these phases are called in lines 6, 7, 9, and 10, respectively. The extraction and normalization phases are called for every sentence (line 4) and for every annotation type (line 5).

After all sentences are processed, underspecified and ambiguous TEs are subject to analysis in the disambiguation phase. For this, all extracted expressions, which are part of other TEs, are removed (e.g. “Tomorrow morning” is tagged as one TEs, instead of two different expressions). In the next step, all remaining TEs are searched for values starting with *UNDEF*. For these expressions, the reference time and the relation to the reference time are determined, and the values are disambiguated according to this information. In the cleaning phase, all invalid TEs are deleted, i.e., expressions identified by negative rules and thus expressions with the value *REMOVE*.

This chapter presented the important basic concepts need to make possible the understanding of other parts of this work, including the temporal tagger used in this research. The next chapter presents the state of the art related to the subject of this dissertation.

3 State-of-the-art

This chapter presents the main works related to the scope of this research. We present papers related to temporal tagging in the clinical domain. The scientific databases we have searched for include ScienceDirect¹, ACM Digital Library², IEEE Xplore³, Springer-Link⁴, Scopus⁵, ACL Anthology⁶ and MEDLINE/PubMed⁷.

In order to achieve the aim of this study, a sample of the words used in the search strings are presented. They are: “temporal tagging”, “time recognition”, “clinical text”, “health document” and “medical narrative”.

The final search strings we used are shown below.

(“temporal” OR “time”) AND (“tagging” OR “annotation” OR “extraction” OR “recognition” OR “information” OR “expression”) AND (“clinical” OR “health” OR “medical”) AND (“text” OR “document” OR “narrative”)

(“temporal tagging” OR “temporal annotation” OR “time extraction” OR “time recognition” OR “temporal information” OR “time tagging” OR “time annotation” OR “temporal extraction” OR “temporal recognition” OR “time information” OR “temporal expression” OR “time expression”) AND (“clinical text” OR “medical text” OR “health text” OR “clinical document” OR “medical document” OR “health document” OR “clinical narrative” OR “medical narrative” OR “health narrative”)

It is important to note that these two similar search strings were used in order to fit the different approach of the scientific databases search engines. The articles were collected up to February 29, 2019.

Our final search string did not have the words “noisy”, “misspelled” and “dirty”, meaning noisy, misspelled or dirty clinical text, because when they were added to the aforementioned scientific databases engines, we did not find articles related to the scope of this research. The only work related to our work is the one by Tissot (2016), which was found by chance in a search at google.com.br⁸. This work was already referred to in the introductory chapter of this dissertation.

According to (KREIMEYER et al., 2017), the majority of Natural Language Pro-

¹ <http://www.sciencedirect.com/>

² <http://dl.acm.org/>

³ <http://ieeexplore.ieee.org/>

⁴ <https://link.springer.com/>

⁵ <https://www.scopus.com/>

⁶ <http://aclweb.org/anthology/>

⁷ <https://www.ncbi.nlm.nih.gov/pubmed/>

⁸ <https://www.google.com.br/>

cessing (NLP) systems for capturing and standardizing unstructured clinical information follow a rule-based approach. The second most prominent approach is a hybrid one. Finally, a machine learning approach alone is the least used strategy to accomplish this task. By hybrid approach we mean a rule-based approach plus a machine learning approach.

Based on the information collected by the same authors (KREIMEYER et al., 2017) from the final set of papers of their study, out of 56 articles, rule-based NLP studies were the most common (33 articles). Hybrid systems were reported in 19 articles. Only 4 articles reported using a machine learning strategy. Temporal tagging in the clinical domain follows the same trend of the bigger research field.

The upcoming sections present the scope of this work, related to the aforementioned approaches. This work dedicates a section to discuss annotated corpora because this kind of input is used in the proposed method (chapter 5), and an annotated corpora was also compiled in this dissertation. The three different approaches to deal with the extraction and normalization of TEs are also discussed. This discussion helped to choose which strategy and which tagger fitted better in the task of extracting and normalizing TEs from noisy clinical text written in Brazilian Portuguese. A special importance is given to rule-based approaches, once our proposed method is based on this type of strategy. The section "machine learning and hybrid approach" discusses alternative strategies that are also possible to be used in order to perform temporal tagging. However, they were not tested, because no other tagger fitted so well as HeidelTime in our requirements. A last section presents the main team results of an important workshop called SemEval, which was dedicated to clinical texts for three years.

3.1 Annotated corpora

Annotated corpora is an important issue in the field of temporal tagging, no matter if the tagging task is done by a rule-based approach, a machine learning approach or a hybrid approach. An annotated corpus is a set of examples which were tagged (manually in most of the cases) and are used as patterns to build up new rules in a rule-based system, to train a machine learning model or to test the effectiveness of both approaches, including a hybrid one.

The most common temporal information domain is the news domain. TempEval (VERHAGEN et al., 2007) and TempEval-2 (VERHAGEN et al., 2010) compiled examples of temporal challenges in the news domain. These challenges were part of SemEval-2007 (AGIRRE; MÀRQUEZ; WICENTOWSKI, 2007) and SemEval-2010 (VERHAGEN et al., 2010) respectively, workshops on semantic evaluation, which are regularly held, usually yearly.

As examples of corpora in the clinical domain, we have the i2b2 2012 corpus

(SUN; RUMSHISKY; UZUNER, 2013) and the Clinical Tempeval corpus (BETHARD et al., 2015).

Most corpora available for temporal information are in English. However, there are already some temporal information corpora built upon other languages. One of these corpora is the work by Costa and Branco (2012). In order to build up a new temporal information corpus a huge work and effort is needed, in the sense of hours spent on the annotation process, and the need for trained people in the area being annotated. To create the temporal information corpus in Portuguese Costa and Branco (2012) adapted the first TempEval challenge English corpus (VERHAGEN et al., 2007) to Portuguese. As a first step in their strategy, authors made a copy of the original annotated corpus in English and removed all TimeML markups from it. Afterwards, the corpus was translated by Google Translator⁹ to Portuguese. Once automatic translation was done, a human translator corrected the mistranslations. The second step ended up with three corpora. The annotated corpus in English (TempeEval challenge corpus), the unannotated copied corpus in English and the unannotated corpus in Portuguese (translated by Google Translator). The annotation of the Portuguese corpus worked in this way. The annotators putted the annotated corpus in English side by side with the unannotated translated corpus in Portuguese. They looked at where the corpus was annotated in the English corpus and annotated the corresponding word or group of words in the Portuguese corpus. For example, if the word "today" was marked and normalized in the English corpus, they marked the word "hoje" (today) and normalized the same corresponding word in the Portuguese corpus. The normalization process was just copied from the English corpus to the Portuguese corpus, since the normalization is always in English, according to the TimeML standard. When the Portuguese corpus was compared with its original source corpus in English, there was a slightly better result in favor of the Portuguese corpus. The comparison was made through some off-the-shelf machine learning algorithms, most of them present in the Weka tool (HALL et al., 2009).

Another example of a clinical corpus in English is the one presented in the work by Viani et al. (2018), who described the creation of a mental health corpus. The authors present how hard was the annotation process to distinguish between DATE and DURATION types. They also compared the I2B2 Corpus and the THYME corpus (STYLER; WILLIAM et al., 2014; IV et al., 2014) with the corpus described in the article. The interest in clinical corpus is a continuous need, even in other languages, like French. Campillos et al. (2018) describe the creation of a clinical corpus in French with many different entities and relations, including temporal expressions, events and temporal relations.

In the next section, we will focus on one of the strategies which take advantage of annotated corpora: the rule-based approach.

⁹ <https://translate.google.com/>

3.2 Rule-based approach

As mentioned previously, the rule-based approach is the most common strategy within the temporal tagging research area. The clinical domain usually presents narratives with particular terms, abbreviations and deviant spelling, which makes the classification task harder when compared to news texts, for example (the news domain is the commonest field studied in temporal tagging tasks). In this case, a rule-based approach presents a more trustworthy result when temporal tagging is the scope. This understanding is presented by (CHANG; MANNING, 2012), when its SUTime system is presented. It is a deterministic rule-based system for temporal tagging recognition of text written in English; the system is part of the Stanford CoreNLP project. It was designed having extensibility in mind, which allows the addition of new rules in a simple way. However, the system was made to cope with news-style text only.

Lee and colleagues (LEE et al., 2014) present a different approach which is similar to a rule-base strategy (the system is called UWTime). Authors explain that “We use a Combinatory Categorical Grammar to construct compositional meaning representations, while considering contextual cues, such as the document creation time and the tense of the governing verb, to compute the final time values.” In this work, authors present a context-dependent semantic parsing approach to learn to identify and interpret time expressions in English (extract and normalize the temporal expressions).

GUTime (VERHAGEN et al., 2005) is another system that is able to deal with temporal expressions. It was one of the first systems to perform extraction and normalization of temporal expressions. It is considered a simple rule-based system able to extract explicit expressions, relative expressions and underspecified expressions. It uses the document creation time as the unique reference to relative and underspecified temporal expressions (STRÖTGEN; GERTZ, 2016).

The current state-of-the-art system for temporal tagging is HeidelTime (STRÖTGEN; GERTZ, 2010). This is also a rule-based system, however it is a project which started from scratch with the purpose of deal with multilingual and multi-domain text (STRÖTGEN; GERTZ, 2015). HeildeTime was made to be able to extract temporal expressions from four different domains: news, narrative, colloquial (short text) and scientific. The most studied domain is the news one, what follows the main stream in the temporal information research field. As HeidelTime was created to extend different languages, it has been already extended to languages such as Spanish (GERTZ, 2013), French (MORICEAU; TANNIER, 2013), Chinese (LI et al., 2014), Croation (SKUKAN; GLAVAŠ; ŠNAJDER, 2014) and Italian (MANFREDI et al., 2014). The Spanish, French, Chinese and Italian mentioned extensions of HeidelTime were made to cope with the news domain. The Croation extension was made to deal with the narrative domain. A peculiar extension of HeidelTime was presented by (HAMON; GRABAR, 2014), which is able to work with

medical clinical narratives in French.

Hao and colleagues (HAO et al., 2018) created a tagger called Temporal Expression ExtractoR (TEER). The system can be considered multilingual, because it was made to deal with clinical texts in two languages, English and Chinese. The system is divided in two parts. In one-part, heuristic rules (regular expressions alike) are created based on the observation of temporal expressions annotations (manually annotated). In the other part of the method, the clinical texts are split into sentences, using the NLTK sentence boundary identifier and some custom rules created to deal with especial cases. After that, a window with 2 or more words and 8 or less words extracts pattern candidates in each sentence from a manually annotated training set. The support and confidence of each extraction is calculated over the original texts. The new pattern is generated, ranked and kept to the final patterns list if, the support is greater than 3 and the confidence is higher than 0.6. Only then, the new pattern is run over the test set. The window size, support and confidence values were empirically found. According to the authors, support is the count of correct matches and confidence is the rate of correct expression matches among all matches. Finally, the heuristic rules and the ranked patterns are joint to extract temporal expressions of newly coming clinical texts. The method also makes the normalization step which follows the tag TIMEX3 of the TimeML guideline. This strategy is inspired in the work of (HAO; RUSANOV; WENG, 2013). According to its authors, TEER presented better results then the most know temporal taggers, including HeidelTime.

As already explained in chapter 2, the temporal tagging task is divided into two steps: extraction and normalization. All systems described above perform both steps. In the next section the machine learning and hybrid approaches are presented.

3.3 Machine learning and hybrid approach

In the context of temporal information recognition, machine learning is understood as a term that groups artificial intelligence techniques such as artificial neural networks, support vector machines (SVM), conditional random fields (CRF), deep learning and so on. The use of these techniques for temporal expression recognition is quite rare because these approaches do not give a good recognition output when it is measured. However, these techniques are commonly used in hybrid approaches which mix machine learning and rule-based systems created to extract and normalize temporal expressions.

Gupta, Joshi and Bhattacharyya (2015) present an example of machine learning strategy to recognize temporal expressions using an artificial neural network. In their system, the classification output labels indicate if the type of the temporal expression is time, date, duration or frequency. Beyond these 4 types, the network is built to check where is the begin, the middle (inside) and the end (outside) of each word classified as

one of the 4 temporal expression types. To train and test the classifier, news and clinical English text corpora was used.

Liu and colleagues (LIU et al., 2019) presented a deep learning way to cope with TEs. The main scope of their article is temporal relation, however, it uses a different approach to extract temporal expressions and events. The article uses a deep learning approach which uses two Long Short-Term Memory (LSTM) layers combined with a Convolutional Neural Network (CNN) layer to extract entities (temporal expressions and events). The mentioned approach considers the context where the entities are inserted and the position of the entities and its neighbors in embedded matrixes and vectors respectively. The dataset used has clinical texts in Chinese. Each text has different sections like “chief complaint” and “history of present illness”, which led the researchers to develop different ways to get different candidate entities. This feature can be seen as a limitation to use the method in other datasets that do not have the same structure.

The most currently used classifier in temporal information extraction and normalization is Conditional Random Fields (CRF), especially because it uses a statistical approach that takes the correlation between words (tokens) into account. UzZaman and Allen (2010) created a hybrid system called TRIPS to extract and normalize temporal expressions that uses CRF in the extraction step. They used a word-by-word classification for temporal expressions which are represented by a B-I-O (which stands for "Beginning of the sequence", "Inside the sequence" and "Outside the sequence") encoding with a set of mostly lexical features. TRIPS was tested over a news domain and a medical domain in English.

Moharasar and Ho (2016) proposed a hybrid system which also uses the CRF classifier. The strategy used in their work starts with HeidelTime extracting temporal expressions from clinical text in English taken from the i2b2 2012 challenge. Along with the HeidelTime extracted temporal expressions, the authors generated lexical features to train a linear chain CRF.

Moharasan and Ho (2019) have continued to propose hybrid systems which use the CRF classifier. In a work about temporal relation, in the temporal extraction section, the authors created specific rules for clinical texts using HeidelTime. Lexical features were created using GENIA tagger. The features extracted by HeidelTime and GENIA were used as input to a CRF classifier. The data used in their work are raw clinical narratives from the i2b2 challenge, in English. The TimeML annotation guideline was followed in the dataset used and in the extraction of TEs by the method proposed by the researchers.

Tang et al. (2013) also worked over the i2b2 2012 challenge clinical corpus. In the temporal expression recognition phase, the authors opted to create a rule-based tool using the programming language Python. Many of the rules created in this tool were adopted from the HeidelTime system. In this study, the CRF classifier was used with a Support

Vector Machine (SVM) to extract events from the clinical text corpus.

As mentioned previously, the hybrid approach is the second most commonly used strategy for temporal information extraction tasks. It loses only to the pure rule-based approach.

3.4 SemEval 2015, 2016 and 2017

The International Workshop on Semantic Evaluation (SemEval) held a task called Clinical TempEval dedicated to clinical texts in years 2015, 2016 and 2017. In all three editions teams were invited to identify temporal expressions, events and temporal relations in texts written in English. In 2015 and 2016, the dataset presented to the participants was about colon cancer. In 2017, the participants were invited to use a training set about colon cancer, but a testing set about brain cancer (domain adaptation). The THYME dataset (IV et al., 2014) was used in these three challenges. The articles with the best results are presented in each version of the event.

SemEval 2015

All teams in the 2015 edition of the event are listed in the work of Bethard and colleagues (BETHARD et al., 2015).

In the Clinical TempEval 2015, Tissot et al. (2015) proposed a rule-based system (HINK) and a SVM approach to identify and normalize temporal expressions. In the same challenge, Velupillai et al. (2015) used a SVM to extract temporal expressions. In the normalization step, a SVM was created for each attribute in combination with an adapted version of HeidelTime. The Begging-inside-outside tagging (BIO) strategy was also applied.

SemEval 2016

The results of the teams that participated in the 2016 edition of the event are listed in the work of Bethard and colleagues (BETHARD et al., 2016).

Cohan, Meurer and Goharian (2016) used in the Clinical TempEval a combination of linear-chain CRF, manually-crafted rules (regular expressions), the BIO tagging strategy and Brown clustering to extract temporal expressions and events. A logistic regression approach combined with rules was used to extract attributes (normalization) of temporal expressions (timexes) and events. In the same challenge Grouin and Moriceau (2016) took advantage of a Conditional Random Field (CRF) and a rule-based tagger (HeidelTime)

to extract and normalize temporal expressions. In this study, the CRF is used to extract timexes, while HeidelTime is used to normalize timexes. The authors proposed two approaches, In the first one, a single CRF is used to extract timexes and events, while in the second approach, a CRF is used to extract each entity (timexes and events). Also in the Clinical TempEval 2016, [Lee et al. \(2016\)](#) used a combination of Hidden-Markov Model and Support Vector Machine (HMM-SVM) with the tagger SUTime to extract and normalize timexes, the study also takes advantage of POS features.

SemEval 2017

All teams related to the 2017 edition of the event are listed in the work of Bethard and colleagues ([BETHARD et al., 2017](#)).

In the Clinical TempEval 2017 [MacAvaney, Cohan and Goharian \(2017\)](#) proposed a linear-chain CRFs approach to extract temporal expression extraction. To identify temporal expression in the text, the Begin, Inside and Outside (BIO) approach was applied. A rule-based strategy was used with more complex temporal expressions. The extraction of the attributes was treated as a classification problem. Each attribute was dealt with by a different CRF model, with the labels of each model corresponded to the possible attribute values. This work was prepared to extract only date and time temporal expression types (excluded the duration and set TimeML types). [Leeuwenberg and Moens \(2017\)](#) used a combination of linear SVMs to extract temporal expressions and events. The authors also used the Stanford POS Tagger for attribute detection. Initially, all tokens with 1,2,3,4,5,6-grams were considered temporal expressions candidates. In this case a gram is equivalent to a token. [Tourille et al. \(2017\)](#) proposed a method divided in two parts. In the preprocessing step, the method uses the cTakes software to extract sentence and token boundaries. They also used HeidelTime in this phase. In the processing step, in order to extract temporal expressions and events, the method uses two LSTMs with a CRF as the last layer. What concerns about the temporal expressions, the study only deals with the extraction step, not the normalization step.

The next chapter presents the methodological procedures designed for this research.

4 Methodological procedures

The aim of this chapter is to explain the methodological procedures, which are the steps taken to create the proposed method. These steps are also important to those who want to reproduce the proposed method, which is described in chapter 5. In this chapter, we assume the training set was already built by the annotators and was also evaluated by the adjudicator. It means the training set is ready to be used in the creation of rules concerning the clinical texts.

Considering the methodological procedures, all TEs of the training set were extracted to become rules or part of the extraction resources of HeidelTime. In this step, TEs were extracted with some context, meaning some words before and after the TE were extracted with each TE. This context was necessary to provide a better understanding of the extracted TEs in the training set. Beyond the context, the whole text where the TE was extracted was also available to be consulted, in case of any doubt appeared about the type of the TE or its normalization.

As the variation of a TE pattern appeared many times in different clinical text of the training set, it was necessary to filter all TEs and select only the unique patterns in order to create rules or resources of extraction at HeidelTime. The filtering process was performed using the Python 3 programming language.

The proposed method comprises three incremental steps, as it is explained in chapter 5. For each step, a different methodological procedure was needed. The "Noisy TEs approach" was the first strategy to be implemented as rules at HeidelTime. It occurred because the unique TEs extracted from the training set did not distinguish between noisy and not noisy TEs.

The "Correct TEs" was the second strategy implemented as rules at HeidelTime. To do so, the rules implemented based on the "Noisy TEs approach" were "cleaned", meaning the part of the rules or resource patterns that were created to deal with "noisy TEs" were removed, remaining only rules and part of the rules or extraction resources that were prepared to cope with "Correct TEs". A manual process was used in this case.

To implement the "N-gram strategy", it was necessary to collect automatically or manually all unique words from the rules or extraction resource files used in the "Correct TEs strategy". The automatic or manual extraction (collect) of the words depended on the type of word, if it was a regular expression or not. If it was a regular expression the manual process was applied, otherwise an automatic process was used. The automatic extraction was accomplished using the Python 3 Language. Once these words were collected, they were used in the "N-gram strategy" that is explained in chapter 5.

Concerning the creation of rules, some details should be explained. The construction of rules had to follow a certain granularity. Considering the text file where rules were created or altered, more granular rules had to be added before the less granular ones, meaning from top to bottom. For example, the more granular rule that matched the TE "03:21 hs" had to come before the less granular rule that matched the TE "09:33" in the rules file. In most of cases HeidelbergTime is able to extract the correct TE. This skill is part of the tagger, once when there is a conflict between rules, HeidelbergTime usually chooses the longest TE as the correct one. However, our experience showed that, in some cases it did not occur, what led us to follow this granularity strategy to prevent extraction mistakes.

Figure 16 shows the whole research structure which was applied in our study.

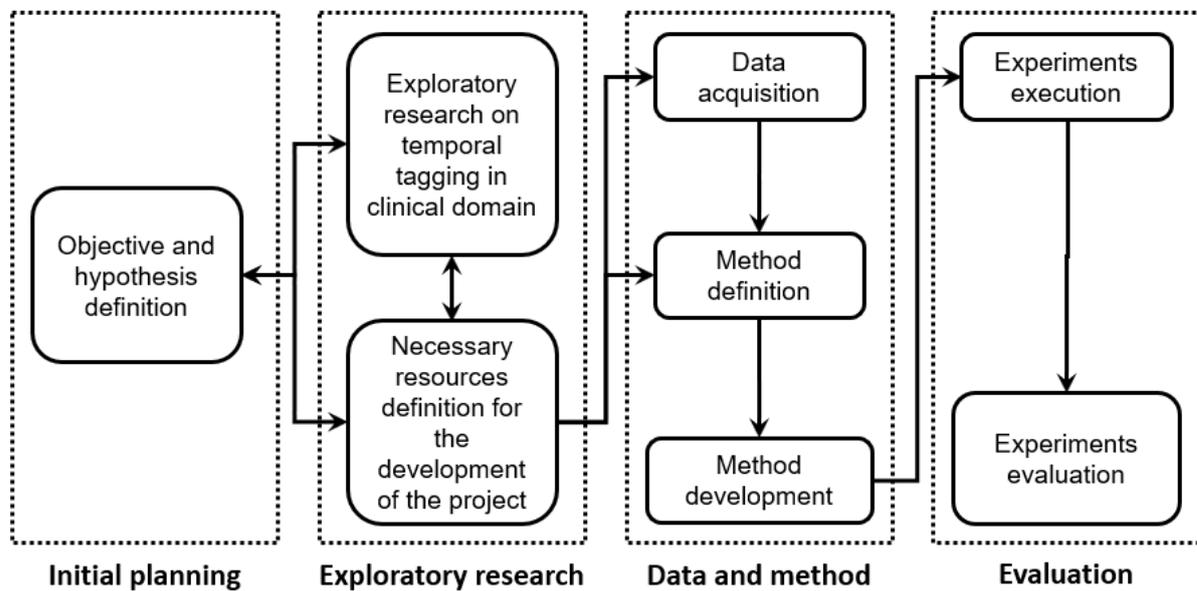


Figure 16 – Research structure developed in this study

The next chapter presents our proposed method.

5 Proposed method

This chapter presents a method to extract and normalize temporal expressions (TEs) in noisy clinical texts written in Brazilian Portuguese using a temporal tagger (AZEVEDO et al., 2018). The method is based on two choices: a rule-based temporal tagger (STRÖTGEN; GERTZ, 2015) and a configuration strategy of a temporal tagger to deal with clinical texts as in Hamon and Grabar (2014).

HeidelTime is a temporal tagger that performs the complete task of temporal tagging, which is the extraction and normalization of TEs. In other words, it is able to identify TEs in a text and normalize them to a TimeML standard. It is a rule-based tagger, according to its authors. The decision to make this type of software to identify time in text instead of a machine learning approach is that a rule-based approach allows normalization to be supervised in a much easier way (STRÖTGEN; GERTZ, 2010). Apart from this characteristic, HeidelTime was also chosen to be the temporal tagger of our research because: it is mentioned as the state-of-the-art tool for the temporal tagging task in many scientific articles, it was built separating the source code and linguistic resources, it has support for Portuguese and for clinical texts. No other tool or technique met these requirements when our study was conducted, whether it was a rule-based approach, a machine learning approach or a hybrid approach.

HeidelTime uses mainly regular expression patterns (pattern resources are explained in section 2.4) to extract TEs and knowledge resources as linguistic clues for the normalization task. HeidelTime was initially designed to be a temporal tagger able to identify TEs written in English and in the news domain only. However, as the project evolved, it was further developed to become a multilingual and domain-sensitive temporal tagger (STRÖTGEN; GERTZ, 2013), (STRÖTGEN, 2015), (STRÖTGEN; GERTZ, 2016). In its current version HeidelTime has been made extensible to other languages, as it separates the source code and language resources like patterns and rules. HeidelTime covers four domains: news, narrative, short text and scientific (clinical). More details about the temporal tagger used in this work are presented in section 2.4.

Hamon and Grabar (2014) presented a form of tuning HeidelTime to identify TEs in clinical text written in English and French. The system is tuned to deal with the mentioned clinical text. The altered HeidelTime rules are trained and tested over the i2b2 2012 challenge corpus (English) and over a French corpus compiled from records of a French hospital. Both corpus have similar size for training and test steps. The work got 0.85 f-measure on the recognition and normalization of TEs in English and around 0.94 in French.

5.1 Method assumptions

The method rests on the following assumptions:

- It has to accept textual data as input;
- It ought to extract and normalize automatically TEs in clinical text written in Portuguese;

5.2 Preprocessing

Clinical texts are as noisy as the routine of the health professionals who write them. Examples of noisy TEs found in clinical texts are: 1) “*sema*” - the standard spelling is “*semana*” (week), 2) Different ways to present an hour and minute - “*05h e 30 min*”, “*15:07*” and “*4h10*”. Thus, they need to go through a preprocessing step.

In the preprocessing step, data is cleaned and prepared. In this step, training and testing sets are changed to lowercase, while some typing errors are corrected. The correction is done in two steps: identification of errors and correction of all occurrences of the identified patterns (errors). Examples of the preprocessing are: “*ago/06*” which was corrected to “*ago/06)*”. The TE “*3 dia(s)*” was corrected to “*3 dia (s)*” and “*59 anos#*” that was corrected to “*59 anos #*”. In all cases a white-space had to be added before special characters like “)”, “(” and “#”. It is necessary to make the tagger able to recognize these TEs. It is a HeidelTime limitation. Beyond this, each clinical text was turned into a single sentence because some TEs with two or more words were separated in two paragraphs, one word at the end and the other word at the beginning of different paragraphs (a line break problem). This problem is shown in table 4

Table 4 – Line break problem

Noisy Temporal Expression	Preprocessed
yesterday	yesterday morning
morning	

The processing step is presented in the next section.

5.3 Processing

The processing step starts by manually transforming each TE of the training set into a new rule and/or a new (pattern and/or normalization) resource in HeidelTime. It is important to highlight that we modified the language resources (Portuguese) part

of HeidelTime only; the algorithmic part (Java code) was not changed. The processing step was done in three different incremental approaches, which supposedly makes the last approach better than the others. In all of them, the already existing language resources in HeidelTime for Portuguese were kept and appended with TE from the clinical texts. When each rule or resource file was created or modified, HeidelTime was executed and tested to check if the rule worked correctly for the case it lead the creation of that rule or of the resource file. Figure 17 illustrates the process and also shows the main contributions of this study, which are represented by parts (b) and (d) of the same figure.

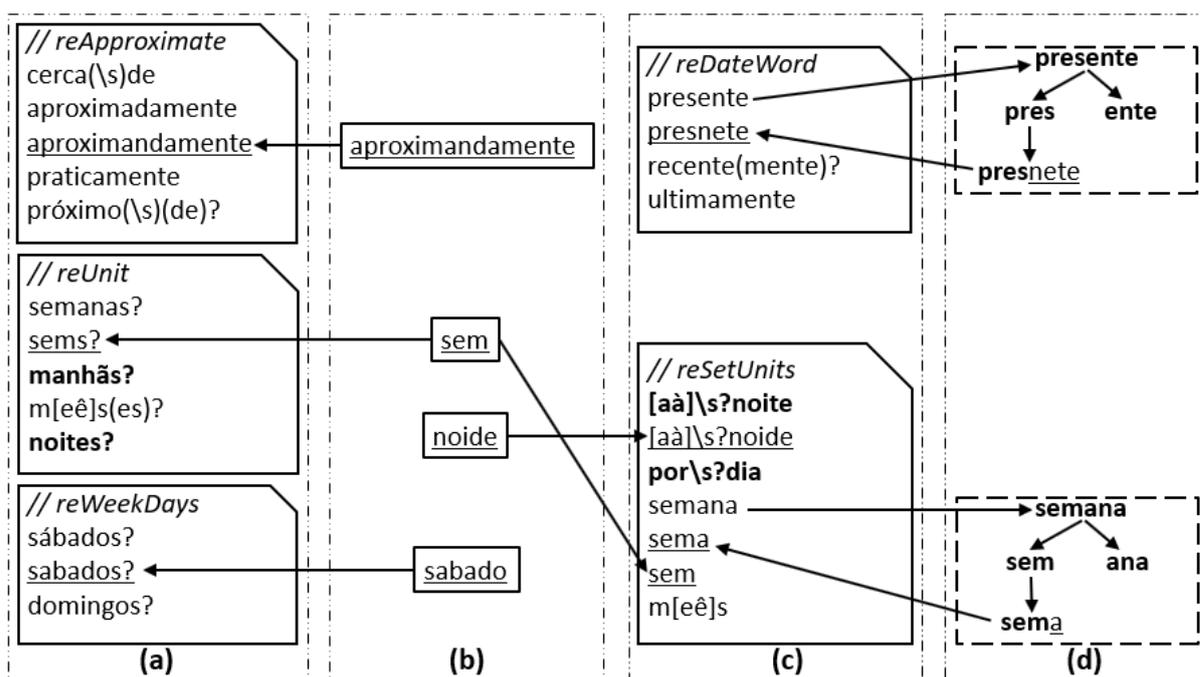


Figure 17 – Method processing step

Item (a) contains three pattern resources called reApproximate, reUnit and reWeekDays. These files already existed in the HeidelTime original version, taken from github¹. Words that follow the name of each file are TEs or part of TEs in that file. Item (c) contains two pattern resources called reDateWord and reSetUnit. The first file also already existed in the original HeidelTime version. However, the second file (reSetUnits) was developed by us, as it was needed in some of HeidelTime rules. Each approach in our method will be explained below.

The first approach is called “Correct TEs”. It is the addition of correctly written TEs found in the training set of the clinical corpus (in bold). They were added in pattern resources that already existed in HeidelTime or were added to it, which are represented in items (a) and (c), respectively. In this approach, white space errors were not dealt with.

¹ <https://github.com/HeidelTime/heideltime>

In the second approach called “Noisy TEs”, the previous approach contribution was kept, but noisy TEs found in the training set were also added to HeidelbergTime. This approach is represented by item (b), which had its TEs added to items (a) and (reSetUnits in c) (underlined TEs in both cases). In this approach, rules were prepared to treat white-space errors, like in “[aà]\s?noite”, where the regular expression “\s?” allows the rule to match a TE with or without a white-space, like in (à noite or ànoite).

In the third approach called “N-gram TEs”, the second approach is appended with the n-gram process. This is done by dividing in the middle each correctly written word of a TE with length greater than five characters (we used a changeable “n” which is equal or greater than three), and each half (n-gram) was searched for in 715 clinical texts not used in the training and test sets. When the evaluated word had an odd number of characters, both possibilities were tested. First, the left half remained with the extra character, like in “próximo” (next), which is divided as “próx” and “imo”. Afterwards, the right half remained with the extra character, thus, “pró” and “ximo”. The aim of this approach is to find misspelled words of TEs in order to add the found patterns to HeidelbergTime rules. The “N-gram TEs” approach is represented in items (c) and (d). Item (d) illustrates the n-gram strategy; thus, the TE part “*presente*” (current) has a character length equal to eight. The word was split in two parts “*pres*” and “*ente*”. Afterwards, each part of the word (n-gram) was searched for in all 715 clinical texts. In item (d), the misspelled TE “*presnete*” represents a finding result of this process. Finally, the noisy TE is added to item (c). This process was done for each correctly written word of TEs present in the pattern resources of the “Correct TEs” approach. The “N-gram TEs” approach is a type of string similarity strategy used to recognize misspelled words. It was adopted in this research as an alternative way to cope with the noisy TEs in clinical texts written in Brazilian Portuguese, once a similar approach was already done based on the Edit Distance algorithm (TISSOT, 2016), a well-known algorithm for this purpose.

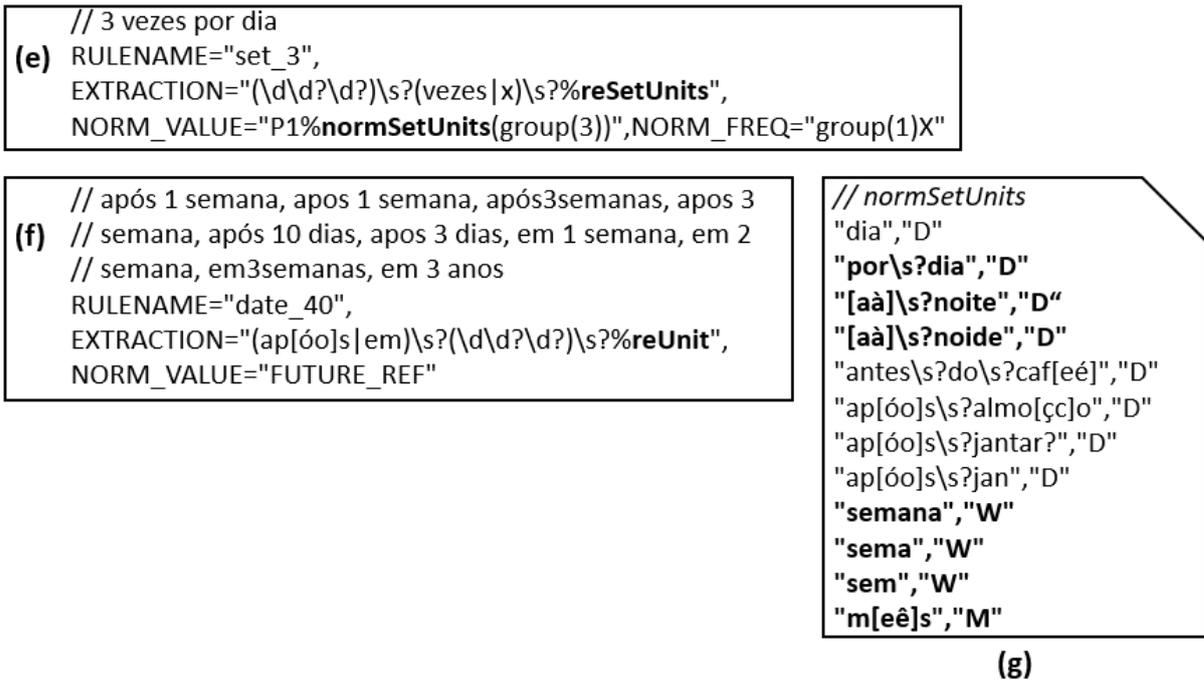


Figure 18 – Example of created rules and a normalization resource

HeidelTime also needs other resources apart from pattern resources to make the whole Temporal Information Task. Items (e) and (f) are examples of rules created by us in the n-gram version of the proposed method (figure 18). Item (g) is an example of a normalization resource also created by us in the same version of the method.

Item (e) is a rule created to deal with TEs of the type SET. This rule intends to cope with TEs like “3 vezes por dia” (thrice a day). Its pattern resource “reSetUnits” is part of the extraction section of the rule. The normalization resource “normSetUnits” is part of the normalization section of the rule.

Item (f) is a rule created to put up with TEs of the type DATE. The rule deals with TEs like “após 1 semana, apos 1 semana (the correct spelling is *após*), após3semanas (the correct spelling is *após 3 semanas*), apos 3 semana (the correct spelling is *após 3 semanas*), após 10 dias, apos 3 dias (the correct spelling is *após*), em 1 semana, em 2 semana (the correct spelling is *semanas*), em3semanas (the correct spelling is *em 3 semanas*), em 3 anos”, which means (after 1 week, after 1 week, after3weeks, after 3 week, after 10 days, after 3 days, in 1 week, in 2 week, in3weeks, in 3 years). Its pattern resource “reUnit” is part of the extraction section of the rule. The normalization resource section of this rule is written within the rule, FUTURE_REF, to indicate an imprecise future date.

Rules created in this study, the TreeTagger^{2,3} and other resources used or created

² <https://www.cis.uni-muenchen.de/schmid/tools/TreeTagger/>

³ <https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html>

in this research are available at:

http://www.ppgia.pucpr.br/~paraiso/pln_em_textos_clinicos/

The next chapter presents the methodological procedures of this dissertation.

6 Dataset annotation

The aim of this chapter is to present the dataset used in this research. It details all steps of its annotation process and presents the guidelines produced in the annotation step.

6.1 The dataset

The dataset used in this study has clinical reports that were compiled from records from hospitals in two different Brazilian states, Paraná and Rio Grande do Sul. Its use was approved by the Research Ethics Committee of PUCPR under number 35849378 of December 2015. Most of these clinical texts are discharge summaries. This dataset has 1025 unannotated clinical reports with TEs. From this total, 310 texts were taken apart to be annotated. In the smallest set, 190 texts were separated to be a training set and the rest, 120 texts, was reserved to be a test set. The size and split strategy of the training and test sets were the same used in the famous i2b2 corpus. The split strategy was approximately 60% for the training set and 40% for the test set. A larger corpus was not annotated because of personnel and deadline limitations. Among the annotated set, the smallest text had 517 characters, while the biggest text had 5,725 characters, including white-spaces.

As temporal expressions were the aim of this study, it is important to say that the start or end date of each clinical report was not available in the texts we received. The presence or absence of this information makes difference for a temporal tagger, because indicating if a reference date will be passed to the tagger, or if the tagger will have to find it automatically in the given text makes difference. In a practical way, if the reference date was available in each clinical text, HeidelTime should be setup to receive news texts, narrative texts or short texts (SMS). However, if the reference date was not available, HeidelTime had to be setup to scientific (clinical) text style. In other words, it is a parameter required by the tagger. It is why the tagger was setup to scientific (clinical) texts.

Concerning the creation of these clinical reports, they were created directly in a computer, meaning non of them was first written in a paper sheet and then, typed in a computational system. First the professional, usually a student or a resident physician, evaluates the patient and then, types her/his report about that patient in the computer. Reports used in our research came from hospitals database, as a Structured Query Language (SQL) file.

6.2 The annotation process

This section describes the annotation process key aspects like: annotators, dataset preparation and annotation itself.

6.2.1 The annotators and the adjudicator

The 310 clinical texts, training and test sets, were annotated by four people. They were three medical students, two female and one male, all in their last year of their undergraduate studies. The fourth annotator was a senior nurse with a master degree in health informatics. They were selected for this task because they already had experience on annotating clinical texts. To judge the work of the annotators, who worked in pairs, there was also an adjudicator, who was a master degree student researching on temporal tagging of noisy clinical texts written in Brazilian Portuguese: the author of this dissertation. His knowledge about the theme, considering the construction of the aforementioned dataset came from some guidelines related to TEs research (FERRO et al., 2001; PUSTEJOVSKY et al., 2003; SAURI et al., 2006; GROUPE et al., 2009). It is important to make clear that the adjudication process was also based in the consensus with the annotators and the professor coordinating the annotation process. The aforementioned professor teaches subjects to students of masters and doctorate in health informatics, she is part of the graduate health informatics program. The adjudication was based on the guideline which is presented in appendices A and B. The guideline was also built up in consensus with the same team involved in the corpus annotation.

6.2.2 The dataset preparation

The dataset was split in 16 files. Each file was made up with 20 clinical texts, except for the last file, which had 10 texts. This first process is shown in figure 19.

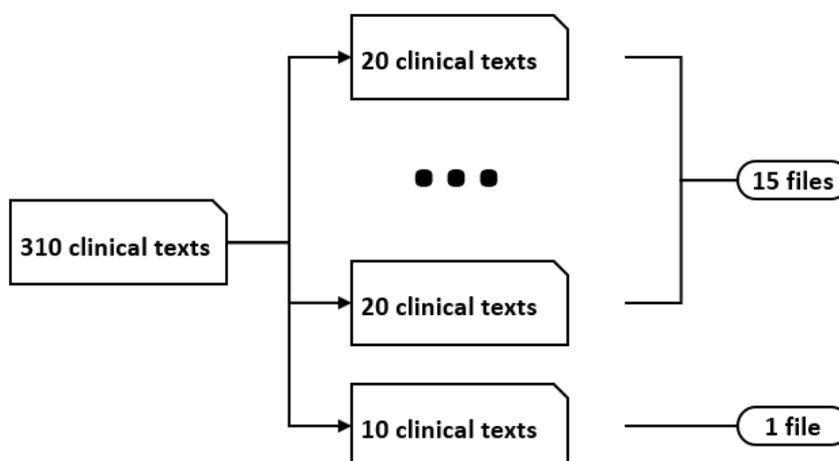


Figure 19 – Files split strategy

To distribute the texts in each file in a balanced way, the clinical texts were sorted in a descending order according to their length. Subsequently, the first file received 10 of the longest clinical texts and 10 of the shortest texts. The original dataset having 310 texts was turned into a smaller one with 290 texts and the process was restarted. So, the second file received 10 of the longest clinical texts and 10 of the shortest texts; this logic was repeated until all texts were distributed in 16 files. The creation of the first three files is illustrated in figure 20. This process was performed to allow annotators to work with different lengths of texts.

All texts sorted from bigger to smaller

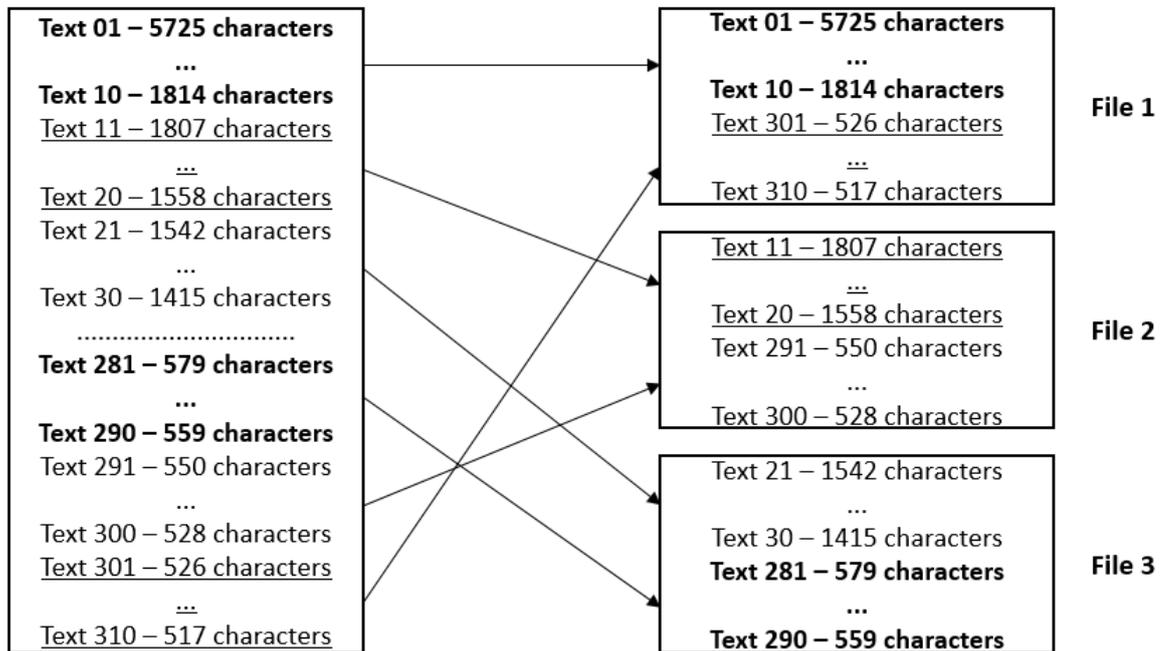


Figure 20 – Files split by clinical text character quantity

Files were organized in a way that each annotator worked in pairs with all other three annotators, as shown in table 5. The table also indicates which file was used as training set, test set and how many clinical texts each file had.

Table 5 – Distribution of files between annotator pairs

File	Train	Test	N° of Clinical Texts	Annotators
01	X		20	annotator 3, annotator 4
02		X	20	annotator 1, annotator 2
03	X		20	annotator 1, annotator 2
04		X	20	annotator 1, annotator 3
05	X		20	annotator 1, annotator 3
06		X	20	annotator 1, annotator 4
07	X		20	annotator 2, annotator 4
08	X		20	annotator 1, annotator 2
09	X		20	annotator 3, annotator 4
10		X	20	annotator 2, annotator 4
11	X		20	annotator 1, annotator 4
12		X	20	annotator 2, annotator 3
13	X		20	annotator 2, annotator 3
14		X	20	annotator 3, annotator 4
15	X		20	annotator 2, annotator 3
16	X		10	annotator 1, annotator 4

Texts assigned for annotators to tag TEs were pre-annotated by HeildeTime with its original rules version, taken from github¹. The tagger was used only to extract TEs. Annotators were warned to extract TEs not extracted by HeildeTime and also to exclude TEs wrongly annotated by the tagger. All TEs were normalized by annotators. Figure 21 presents a sample of pre-annotated TEs (extracted) by HeildeTime. No reference date was passed to the tagger as a parameter, once it did not exist in the dataset used in this study.

```

<TIMEX3 id="T0" type="DATE" text="abril 2007" spans="145~155"/>
<TIMEX3 id="T2" type="DATE" text="1 dia" spans="576~581"/>
<TIMEX3 id="T3" type="DATE" text="06/06/2007" spans="817~827"/>
<TIMEX3 id="T7" type="DATE" text="2001" spans="1785~1789"/>
<TIMEX3 id="T8" type="DATE" text="03/09/04" spans="1992~2000"/>
<TIMEX3 id="T11" type="DATE" text="ao dia" spans="2695~2701"/>
<TIMEX3 id="T12" type="DATE" text="8 horas" spans="2728~2735"/>
<TIMEX3 id="T13" type="DATE" text="há 20 anos" spans="3245~3255"/>
<TIMEX3 id="T14" type="DATE" text="há 3 anos" spans="3316~3325"/>
<TIMEX3 id="T15" type="DATE" text="atualmente" spans="3445~3455"/>

```

Figure 21 – Pre-annotation samples

It is important to explain that the type DATE in figure 21 is not correct, as explained, the aim of the pre-annotation was to extract TEs. The normalization had to

¹ <https://github.com/HeildeTime/heildetime>

be carried out by annotators. In other words, annotators received all pre-annotated TEs marked as DATE.

Regarding the pre-annotation, the MAE tool^{2 3} (STUBBS, 2011; RIM, 2016) was the software used in the annotation process. However, as the pre-annotation was made by HeidelbergTime and its output file is not compatible with the input file of the MAE tool, it was necessary to develop a parser to convert the HeidelbergTime output file into a compatible input file to the MAE tool. The parser was developed by us using the Python 3 programming language.

The next section explains how the annotation was carried out.

6.2.3 The annotation

The annotation process comprised two steps. The annotators training was the first step. The annotation of the definitive dataset and its adjudication was the second step.

6.2.3.1 Annotators training

The annotators training was carried out through three pilot tests. Before the beginning of tests, the adjudicator prepared some information based on guidelines he had studied and sent to annotators to be read. This was done to make them familiar with some temporal tagging concepts like TEs types, ways of extracting and normalizing TEs etc.

The first pilot test was a face-to-face meeting, when annotators got used to MAE tool (figure 22). At this meeting, each pair of annotators tagged the same two clinical texts, as a way to encourage ideas and answer doubts. When the meeting ended, annotators took six texts to be tagged at home. Pairs of annotators were not static, meaning each annotator was part of a pair with different annotators. This arrangement was made in the file they received with the clinical texts.

² <http://keighrim.github.io/mae-annotation/>

³ <https://github.com/keighrim/mae-annotation>

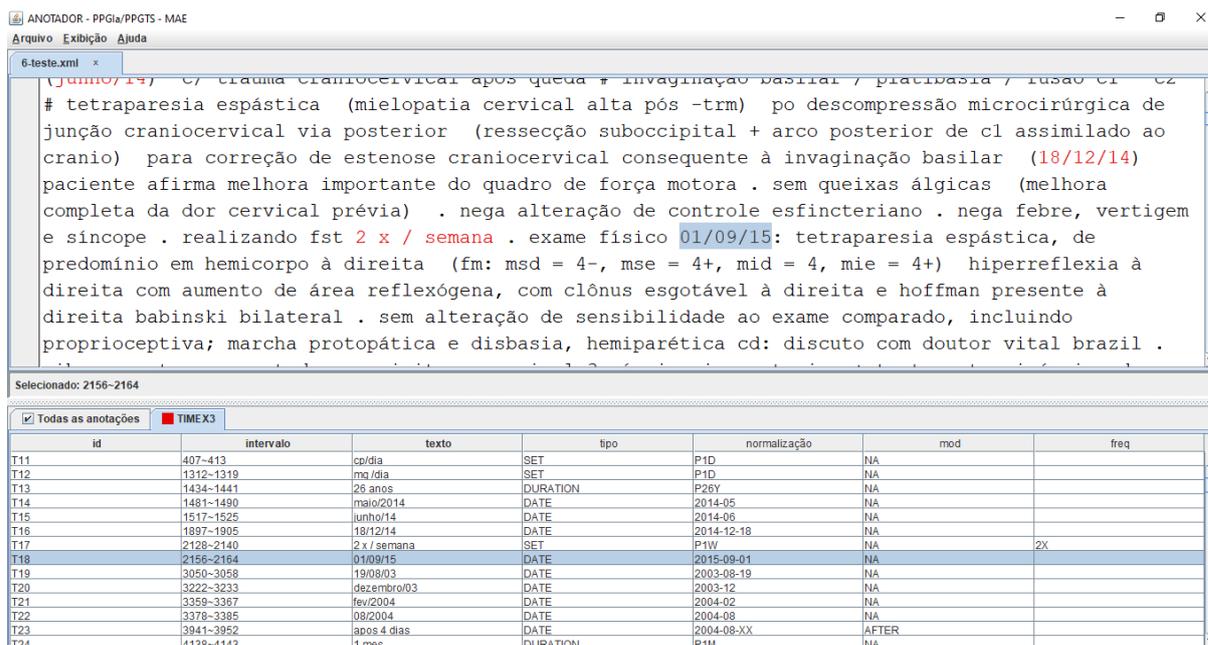


Figure 22 – Screenshot of the MAE tool custom version used by annotators

When all texts were tagged, the adjudicator analyzed the work of each annotator placing some correct samples and all incorrect TEs samples extracted and normalized in a spreadsheet. The adjudicator corrected all incorrect tagged TEs and also explained what was wrong in a column called "details". This spreadsheet collected evaluation results by all annotators and was shared with them, as a guide where they could have a quick look when doubts appeared in further pilot tests.

They could thus learn from hits and mistakes of other annotators. To facilitate the communication among the team members, a work group was created in a popular instant message mobile application. This group was operational during the whole annotation process: annotators training, final annotation and adjudication. This group was the channel to solve doubts in the adjudication process, being the place where the consensus was practiced when ambiguous or confusing TEs appeared.

The second pilot test was similar to the first. When the adjudicator analyzed the six clinical texts the annotators tagged at home, the formerly used spreadsheet was appended with new correct and incorrect samples extracted and normalized TEs found in the annotated texts of the second test. Again, each mistake was explained and the spreadsheet was shared with all annotators.

The third and last pilot test did not include a face-to-face meeting. This time, annotators worked in 10 new clinical texts. The adjudicator analyzed the correct and incorrect annotations and appended them to the same spreadsheet. To minimize doubts by annotators, the adjudicator had individual online meetings with each annotator, where

they discussed each incorrect tagged TE.

It is worth mentioning that three annotators had been previously part in other annotation projects concerning clinical texts. However, the annotation of TEs was a novel experience to all of them. The choice of TEs which would be annotated in this dissertation was made in a practical way. In other words, only TEs that were considered useful on the daily routine of healthcare professionals, especially physicians and nurses, were annotated. This process is better exemplified in the guideline, especially in sections A.12 (Portuguese) and B.12 (English).

It is also important to say that Cohen's Kappa increased considerably between the first and the third pilot test. During the three tests, the guideline adopted in the final version of the dataset was drafted, being amended and improved after each pilot study. The guideline can be found in appendices A and B.

6.2.3.2 Final annotation

The annotation of the final dataset brought about very few doubts. This was due to the fact that almost all annotators doubts were solved by the guideline (appendices A and B) to the process of negotiation (consensus) during the pilot tests. The few questions that appeared were quickly solved in the instant message group already mentioned.

6.2.3.3 Adjudication

The adjudication process was also carried out in the MAE tool (figure 23). It was more complex than the construction of the final dataset by the annotators. The complexity is shown in various cases that follow.

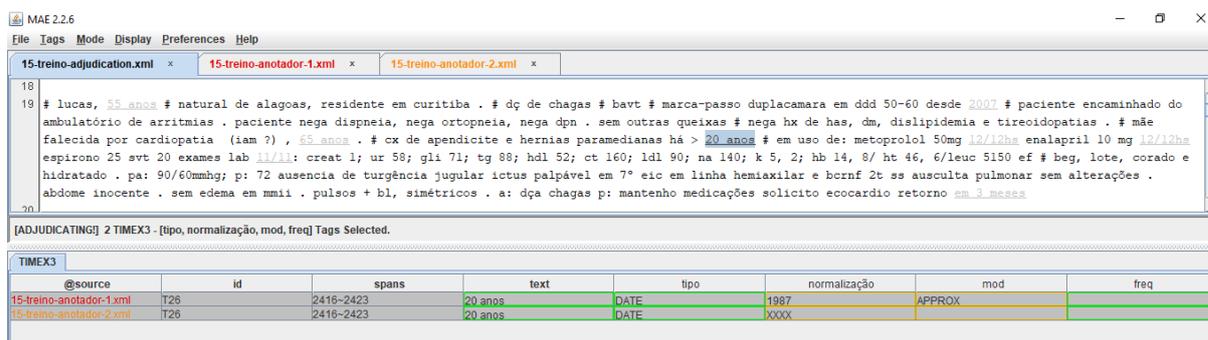


Figure 23 – Screenshot of the MAE tool version used by the adjudicator

It is important to note that all dates follow the Brazilian date pattern which is from left to right: **day[separator]month[separator]year**. Examples are: 13/02/1983, 25 de março/95, 25.07, 30-11-01, 23 02 83, 12|2007 and any other separator.

1. “paciente portados de has, mutação do gene da protrombina em anticoagulação crônica com warfarin com iam e avc em 2001 com sequela (déficit motor à esquerda) com cat de 2001 evidenciando hipertrofia de ve e coronárias normais . ergometria de **06/04** com baixa capacidade funcional . em 30/12/04 iniciou com dor retroesternal, em aperto, intensa sem irradiação”

Details:

- Both annotators tagged in the same way.
 - Annotator 1 normalized as: 2004-04-06
 - Annotator 3 normalized as: 2004-06
 - It is important to note the variability of the date patterns that the same person uses in the same text (perhaps more than one person writing the text). The representation used is 06/04 and 30/12/04 in the same text.
 - The annotator 3 choice was kept in the final version of the dataset. It seems to make more sense, because 06/04 is chronologically written before 30/12/04. It also seems that the day when the exam was made is not so important like the month it occurred. However, people could easily disagree. This is an example of a not easy choice to take, because 06/04 can mean **day/month** and also **month/year**.
2. "paciente com: - dm 2 . - obesidade . - has . - dislipidemia . - cardiopatia isquemica (stent adae 1/3 proximal 03/2004, stent amgcx 1/3 distal 09/2004) . apresentando angina, realizou ergometria positiva, tendo feito cateterismo dia **11/02** que demonstrou reestenose grave intra-stent da amgcx, internou para realização de angioplastia intra-stent, com sucesso angiográfico e fluxo timi 3, sem implante de stent ."

Details:

- Both annotators tagged in the same way.
 - Annotator 1 normalized as: 2004-02-11
 - Annotator 3 normalized as: XXXX-02-11
 - If we consider references to 2004, we can assume that the year is 2004. However, if the text was written in a chronological order, as the writing order of the dates 03/2004 and 09/2004 points out, we can understand that the year of the date is 2005 (2005-02-11). For lack of accuracy the adjudicator kept the option of the annotator 2, because it is the safest, but less accurate.
3. “id: lucas . hma: paciente relata ardência em esofago ao se alimentar, disfagia e odinofagia, vomitos logo apos as refeições, inclusive sem poder beber água desde entrar na dieta solida (**há praticamente 1 mês**) .”

Details:

- Both annotators tagged the TE in the same way.
 - Annotator 1 normalized it as: DURATION P1M
 - Annotator 3 normalized it as: DATE 2016-05
 - Annotators often got confused in cases like this, where a TE could be normalized as DURATION or DATE, depending on the interpretation of the context.
 - As agreed in the guideline, annotator 3's choice was kept in the final version of the dataset, because the phrase starts with the word "há", indicating a past TE. However, annotator 1's choice also makes sense according the context.
4. “**18/02/1106h** - procura direta, acompanhado do pai, em cadeira de rodas . torporoso sem resposta ao estímulo verbal . vomito alimentar em grande quantidade . puncionado acesso . realizado contenção / risco de queda . pai relata que o mesmo ja esteve internado neste hospital em **junho** com o mesmo quadro .”

Details:

- Annotator 1 extracted as: 18/02/1106h
 - Annotator 1 normalized as: 2011-02-18T06:00 - TIME
 - Annotator 3 extracted as: 18/02/110
 - Annotator 3 normalized as: XXXX-02-18 – DATE
 - Annotator 3 extracted as: 6h
 - Annotator 3 normalized as: XXXX-02-18T6 – TIME
 - Annotator 1 and annotator 3 extracted as: junho
 - Annotator 1 normalized as: 2010-06
 - Annotator 3 normalized as: XXXX-06
 - It is interesting to observe how difficult it was for annotators to tag this noisy TE. It is interesting to note that normalization reflects the understanding of the TEs by each annotator.
 - Annotator 1's choice was kept in the ground truth version. Although, it was not an easy decision.
5. "lucas, 63 anos # po tx cardíaco 06/10/2015 / miocardiopatia chagásica # irc agudizada no pós-operatório com necessidade de hd - com melhora após # cateterismo hemodinamico em 2014 - psap 44 mmhg # marca-passo **há 7 anos** // has // dislipidemia // ex-tabagista - parou há 7 anos - 19 maços/ano //"

Details:

- Annotator 2 and annotator 4 extracted the TE as: há 7 anos
 - Annotator 2 normalized as: DURATION – P7Y
 - Annotator 4 normalized as: DATE – XXXX-XX-XX
 - Confusion between DURATION and DATE appeared many times during the annotation process, because, depending on the interpretation of the annotator and the context, both types can be considered correct.
 - Annotator 2's normalization was kept in the ground truth version.
6. "lucas, 63 anos # po tx cardíaco 06/10/2015 / miocardiopatia chagásica # irc agudizada no pós-operatório com necessidade de hd - com melhora após # cateterismo hemodinamico em 2014 - psap 44 mmhg # marca-passo há 7 anos // has // dislipidemia // ex-tabagista - parou **há 7 anos** - 19 maços/ano //

Details:

- Annotator 2 and annotator 4 extracted the TE as: há 7 anos
- Annotator 2 normalized the TE as: DATE – 2008
- Annotator 4 normalized the TE as: DATE – XXXX-XX-XX
- The TE is related to the date when the patient quit smoking. Annotator 2 normalized as 2008, because assumed that it was 7 years before 2015, a explicit date in the clinical report. However, the annotator 4 was not sure about the relation between the two dates.
- Annotator 2 normalization was kept in the ground truth version.

Both explanations above (5 and 6) mention the same text. It is important to note that the TE explained in both cases is the same **há 7 anos**.

"... psap 44 mmhg # marca-passo **há 7 anos** // has // ..."

"... parou **há 7 anos** - 19 maços/ano ..."

However, its type is different, according to the context in which it is used in the corpus.

7. "lucas, 57 anos . # iam ântero-lateral **02/02**: sk sem sucesso, disfunção ventricular e angina pós-infarto . # 2 iams prévios . # has . # dpoc . # dm2 . # ic (fe 49% agosto/06) . # tabagista . paciente interna para realização de cateterismo cardíaco pós iam lateral no dia **02/02** ."

Details:

- In both cases, both annotators extracted the TE as: 02/02
- Annotator 2 normalized both as: 2002-02

- Annotator 3 normalized both as: XXXX-02-02
- This type of TE can be confusing "02/02", because without taking the context into account, it can mean 2nd of February and also February/2002.
- Considering the context, in this case, annotator 3 normalization was correct.

8. paciente refere que **há 6 meses** vem apresentando episódios de dor torácica em aperto

Details:

- Both annotators extracted as: há 6 meses
- Annotator 2 normalized both as: DURATION – P6M
- Annotator 3 normalized both as: DATE – 2014-10
- The type and normalization of TE is confusing. Both interpretations seem to make sense.
- As only one interpretation must be chosen, in the adjudication process, annotator 2 normalization was kept in the final version of the corpus.

9. "tabagista **há mais de 40 anos** (10 cigarros /dia) . # em uso de insulina nph 26ui / 10ui, metformina 850mg café / almoço / janta, aradois 50mg /dia, anlodipina 5mg 12/12 hs, enalapril 20mg 12/12 hs, sinvastatina 80mg, selozok 50mg /dia, aas 100mg, s # nega precordialgia, nega dispneia, ortopneia, dpn . refere claudicação intermitente para moderados esforços . o # beg, lote, corado, hidratado, eupneico, afebril, pa 150/80, fc 74, cp: sp, acv: bcrnf 2t ss, ap: mv+ reduzido difusamente, sem ra, abd: sp, mmii: pulsos reduzidos bilat, sem edema ou empastamento, lab **13/01/14**: glicose 304; glicose pós-prandial 309; ct 119; hdl 21; tg 214; tgo 19; hba1c 6, 70; cpk 72; cr 1, 00; k 4, 7; ur 30; microalbuminuria 24 hs (2114 mg) ; pu (glicose ++; leuc 2) . a # dac, has, dm . p # aumento selozok para 100mg /dia . demais mantido"

Details:

- Both annotators extracted as: há mais de 40 anos
- Annotator 2 normalized both as: DURATION – P40Y
- Annotator 3 normalized both as: DATE – 1974
- In this case, we can assume that annotator 3 is correct, based on the information that laboratory tests (13/01/14) have a certain date validity, so, it is possible to estimate that the year when the clinical report was written was 2014. This shows how hard it is for annotators to identify a correct date, considering interpretation issues.

- Annotator 3's normalization was added to the final version of the corpus.
10. "paciente veio acompanhado da mãe . refere melhora do tremor de mmss com a retirada do haldol, no entanto refere piora da acatisia e discinesia com a introdução da risperidona 02 mg (**0-0-1**) ."
- Extraction: 0-0-1
 - Normalization: SET – P1D
 - Among 4 annotators, only 2 of them, the most experienced ones, knew this temporal expression pattern, which mean: [medicine_quantity_in_the_morning-medicine_quantity_at_lunchtime-medicine_quantity_in_the_evening], **0-0-1** means one medicine at night. While **2-0-1** means 2 tablets in the morning and 1 in the evening.

The adjudication process brought insights about temporal tagging of real world clinical texts, including noisy ones. The average Inter-Annotator Agreement achieved in the construction of this dataset was 0.9018 for the extraction step and 0.7271 for the normalization step (Cohen Kappa). The extraction was calculated following the relaxed matching strategy. The normalization was calculated following the strict matching strategy, considering if the attributes type and value of the TIMEX3 tag were equal for both annotators. However, as there was a high disagreement among the annotators between DATE and DURATION types, if one annotator indicated the type equal to DATE and the other annotator indicated the type equal to DURATION or vice-versa, we also calculated the normalization as correct, which resulted in a normalization with a Cohen Kappa of 0.8922. The mod attribute was not considered, because in most of the cases, only one annotator annotated this attribute in a proper way. The approximate type distribution of TEs in the created corpus is presented in figure 24.

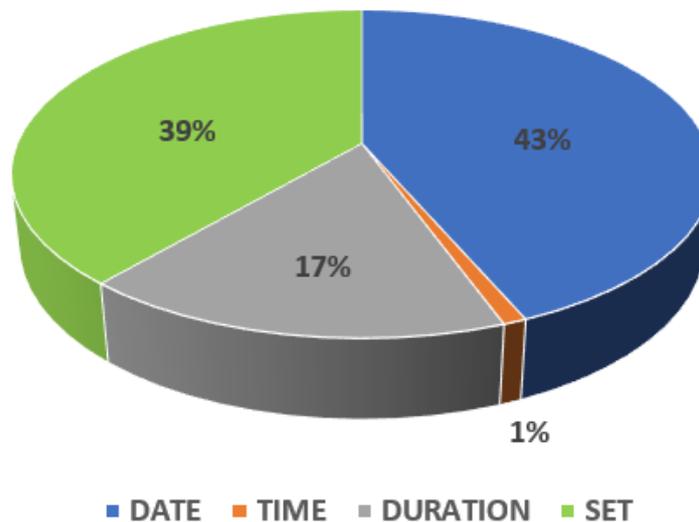


Figure 24 – Approximate percentage of TEs type in the created corpus

Temporal expressions like **2x/dia** (twice a day) can be interpreted as SET type, because it repeats a certain times in a day. However, it can also be understood as DURATION type, considering the first and last time that the medicine is taken during the day, in terms of the lasting of the medicine effect in the person's body. Nevertheless, the second interpretation is not usual, as the first interpretation is considered the most useful one by nurses and physicians, based on the consensus of the annotation team.

TE **mg 12/12** is an example of the recurrent need to have access to additional information in the tagging process. If the SET type TE mentioned was tagged as only **12/12**, it could mean 12/December or December/2012. If we take human understanding of the TE into consideration, it is important to have the "**mg**" part added. If a person is asked to normalize TE **12/12**, three readings could be possible, 12/December, December/2012 or take a medicine every 12 hours. This is why we decided to keep the text particle **mg** in the case of SET type. Although, thinking about how a tagger could handle this problem, a rule-based or a machine learning approach would behave in a different way. For a rule-based approach, the measurement unit **mg** does make difference. However, a machine learning algorithm like CRF would not need the **mg** unit to tag in a proper manner. Considering the human understanding of the TE, we decided to keep the particle **mg**. Other examples of this case adopted in the adjudication process were: 12/12hs, 12/12h, 12/12hrs, 12/12 hs and 12/12 horas. Nonetheless, it is important to say that there were cases where **12/12** meant taking the medicine twice a day and there was no other "word" that helped to distinguish the TE was from a DATE type. In this case only a specific context could indicate that the TE is of the type SET, instead of DATE. An example is "em uso de: - enalapril 10 - **12/12**; - carvedilol 6, 25 - **12/12**; - aas 100;". Around the TE, meaning one or more words before or after the TE are not a pattern that can be tagged, in order to help a human understand that the TE **12/12** is not of the DATE type. The same challenge was found with the TE (thrice a day), which had many forms like: 3x, 3x/d, 3x/dia, 3 x/dia, 3x por dia, 3x / dia.

Another confusing TE is the one that indicates how often a tablet must be taken. Examples of this TE are: 1 cp/dia (one tablet a day), 3 cp/dia (three tablets a day) etc. There are cases where the TE is unambiguous, like in table 6.

Table 6 – Unambiguous examples of SET type

TE	Example	Comment
1 cp/dia seg - sab	marevan 1 cp/dia seg - sab para no alvo sic	1 tablet from monday to saturday
1 cp/dia	acarbose 1 cp/dia	1 tablet per day
1cp ao dia	em uso de: aas 100 mg - 1cp ao dia	
1 cp de 12/12h	enalapril 10 mg - 1 cp de 12/12h	1 tablet each 12 hours
1 cp ao dia	sinvastatina 40 mg - 1 cp ao dia	1 tablet per day
1cp pela manhã	hidroclorotiazida 1cp pela manhã	1 tablet per day in the morning
1cp no café	januvia 100mg 1cp no café	1 tablet per day at breakfast
1 cp contínuo	aas 100 mg 1 cp contínuo	1 tablet per day
3 cp /dia - manhã, tarde e a noite	metformina (3 cp /dia - manhã, tarde e a noite - 850 mg)	3 tablets per day 1 in the morning 1 in the afternoon 1 in the evening
(1 cp/dia - antes do almoço)	glicazida 30mg (1 cp/dia - antes do almoço)	1 tablet per day at lunchtime
2 cp no almoço	glicazida 30mg 2 cp no almoço	Take 2 tablets at lunchtime
1/2 cp /dia	digoxina 1/2 cp /dia	Take half tablet per day
2cp a noite	sinvastatina 20mg 2cp a noite	Take 2 tablets at night

Nevertheless, there are cases where this type of TE is confusing and ambiguous, like in table 7.

Table 7 – Confusing and ambiguous examples of SET type

TE	Example	Comment
4cp / dia	atovastatina 20 mg 4cp / dia	Take all tablets at once or divide them along the day (1 tablet each 6 hours)?
3 cp/dia	metformina 850mg 3 cp/dia	Take all tablets at once or divide them along the day (1 tablet each 8 hours)?
2 cp ao dia	selozok 50 mg - 2 cp ao dia	Take all tablets at once or divide them along the day (1 tablet each 12 hours)?

It is important to explain that the examples in the column "Example" are exhaustive, meaning the example contains the longest length and the maximum context that helps the better understanding of the TE. In other words, there is no extra information available that helps its better understanding.

6.3 The guideline

The guideline built and used during the extraction and normalization of the aforementioned dataset is presented in appendices A and B. It was built based on previous guidelines like TimeML (PUSTEJOVSKY et al., 2003; SAURI et al., 2006) and THYME (STYLER; WILLIAM et al., 2014), especially trying to follow the recommendations by Tissot and colleagues (TISSOT et al., 2015). These recommendations are: make clear what should and what should not be annotated, give as many examples as possible in the guidelines and have as many annotators as possible.

The next chapter presents the experiments and results achieved by the proposed method in this research.

7 Experiments and results

This chapter presents the experiments and results carried out in this dissertation to evaluate our proposed method. The experiments follow the work of Azevedo and Colleagues (AZEVEDO et al., 2018). They were grouped into four experiments. Results were evaluated with three metrics: Precision, Recall and F1-Score. The evaluation was made using the "relaxed matching approach", explained in section 2.3. The normalization step was done only with the True Positive TEs tagged in the extraction step. All experiments were tested with the same 120 clinical texts of the test set already mentioned. Experiments were all done using the scientific domain setup available in HeidelTime (no reference date was informed to the tagger). HeidelTime already has rules to tag and normalize general purpose Portuguese, it means that it can tag some basic TEs in Portuguese. These free of context, general and basic rules were used and incremented in all different experiments, when it was possible.

The first experiment, called "Baseline", used HeidelTime as it was downloaded from GitHub¹. It means the extraction and normalization was done only with the rules and patterns that already existed for the general purpose Portuguese that already existed in the temporal tagger.

In the second experiment, called "Correct TEs", rules and patterns were created and appended to the ones that already existed for Portuguese on HeidelTime. These new rules and patterns were created using the correctly written TEs found in the training set. This experiment implements the first approach of the proposed method, called "Correct TEs".

In the third experiment, called "Noisy TEs", rules and patterns were created and appended to the ones of the second experiment. The new rules and patterns of this experiment were created with the noisy TEs found in the training set. This experiment implements the second approach of the proposed method, called "Noisy TEs".

In the last experiment, called "N-gram TEs", rules and patterns were created and appended to the ones of the third experiment. The new rules and patterns of this fourth experiment were created with the noisy TEs identified by the implementation of the third approach of the proposed method 5.3, which is called "N-gram TEs" approach.

An incremental approach was used in experiments two, three and four. The second experiment incremented the rules and patterns that already existed in the first experiment. The third experiment incremented the rules and patterns that already existed in the second experiment. And following the same logic, the fourth experiment incremented the

¹ <https://github.com/HeidelTime/heideltime>

rules and patterns that already existed in the third experiment.

Some statistics about the altered and created rules are important to have. The complexity of rules is not the same, some are more complex, while others are much simpler. In many cases, the explainability of the rule was favoured, instead of its wide coverage of possibilities. Meaning, one rule can cover many TEs, while another rule, covers only few TEs or just one TE pattern. Tables 8 and 9 show the quantity of rules altered or created (new) in HeidelTime respectively, considering all strategy used in the proposed method.

Table 8 – Quantity of altered rules in all experiments

TEs Type	Quantity of Rules Altered		
	Correct TEs	Noisy TEs	N-gram TEs
DATE	2	2	3
TIME	0	0	2
DURATION	4	4	6
SET	0	0	1

Table 9 – Quantity of altered rules in all experiments

TEs Type	Quantity of New Rules (Created)		
	Correct TEs	Noisy TEs	N-gram TEs
DATE	45	45	45
TIME	7	7	7
DURATION	5	5	5
SET	64	64	64

The quantity of rules in each experiment did not change a lot, they are mostly the same rules in all experiments. However, the content of the rules changed from one approach to the other, specially between the "Correct TEs" and the "Noisy TEs". Changes relate to whether the TE is "Correct" or "Noisy", in other words, if it contains a white-space that should be there or not, for instance. Beside the changes in rules, there were also changes in the content of the resource files. Examples of rules and resource files in different experiments are presented below:

```
// Example a1
// Correct TEs strategy version
// 10/10, 12-08, 10/1, 12-1, 01/07, 1-07, 10-1
RULENAME="date_rafa1",
EXTRACTION="( %reDayNumber[-\/]%reMonthNumber)",
NORM_VALUE="XXXX-%normMonth(group(3))-group(2)"

// Example a2
```

```
// Noisy TEs and N-gram strategy versions
// 10/10, 12-08, 31.7, 10 / 10, 12 - 08, 31 . 7,
// 10 / 1, 12-1, 1.1, 10 / 10,
// 12 - 08, 1 . 07, 08 . 10, 13 . 10, 19 . 10, 28 . 10, 31 . 10
RULENAME="date_rafa1",
EXTRACTION="(reDayNumber\s?[-\./]\s?reMonthNumber)",
NORM_VALUE="XXXX-%normMonth(group(3))-group(2)"
```

```
////////////////////////////////////
```

```
// Example b1
// Correct TEs strategy version
// há dias, há dia
RULENAME="date_rafa45",
EXTRACTION="(há)\s(dias?)",
NORM_VALUE="TPZ-0000-00-02",
NORM_MOD="APPROX"
```

```
// Example b2
// Noisy TEs and N-gram strategy versions
// há dias, ha dia, à dias, á dia,
// hádias, hadia, âdias, ádias, hâdias, hà dias
RULENAME="date_rafa45",
EXTRACTION="(há|hâ|ha|à|á)\s?(dias?)",
NORM_VALUE="TPZ-0000-00-02",
NORM_MOD="APPROX"
```

```
////////////////////////////////////
```

```
// Example c1
// Correct TEs strategy version
// há anos, ha anos
RULENAME="date_rafa43",
EXTRACTION="(há)\s(anos?)",
NORM_VALUE="TPZ-0002",
NORM_MOD="APPROX"
```

```
// Example c2
```

```
// Noisy TEs and N-gram strategy versions
// há anos, ha ano, à anos, á ano,
// háanos, haano, àanos, áano, hãanos, hà ano
RULENAME="date_rafa43",
EXTRACTION="(há|hà|ha|à|á)\s?(anos?)",
NORM_VALUE="TPZ-0002",
NORM_MOD="APPROX"
```

```
////////////////////////////////////
```

```
// Example d1
// Correct TEs strategy version
// às 12h
RULENAME="time_rafa6",
EXTRACTION="(à)s\s%reTimeHour\s(horas|hora|hr|hs|h)",
NORM_VALUE="group(2):00",
NORM_MOD="APPROX"
```

```
// Example d2
// Noisy TEs and N-gram strategy versions
// às 12h, às12h, as 12h, ás12h
RULENAME="time_rafa6",
EXTRACTION="([àáa])?s\s?%reTimeHour\s?(horas|hora|hrs?|hs|h)",
NORM_VALUE="group(2):00",
NORM_MOD="APPROX"
```

```
////////////////////////////////////
```

```
// Example e1
// Correct TEs strategy version
// noite inteira, tarde inteira, manhã inteira
RULENAME="duration_rafa2",
EXTRACTION="(noite|manhã|tarde) inteira",
NORM_VALUE="PT8H"
```

```
// Example e2
// Noisy TEs strategy version
// noite inteira, tarde inteira, manhã inteira
```

```
RULENAME="duration_rafa2",
EXTRACTION="(noite|manh[ãa]|tarde)\s?inteira",
NORM_VALUE="PT8H"
```

```
// Example e3
// N-gram strategy version
// noite inteira, tarde inteira, manhã inteira,
// noide inteira, noideinteira
// noide (noite - night) was a noisy part of the TE
RULENAME="duration_rafa2",
EXTRACTION="(noi.e|manh[ãa]|tarde)\s?inteira",
NORM_VALUE="PT8H"
```

```
////////////////////////////////////
```

```
// Example f1
// Correct TEs strategy version
// duração de 1 hora, duração de 4 horas, em cerca de 12 hora,
// cerca de 12 horas
RULENAME="duration_rafa5",
EXTRACTION="(duração|cerca|evolução|por)\s(de)\s(\d\d?\d?)\s(horas?)",
NORM_VALUE="PTgroup(3)H"
```

```
// Example f2
// Noisy TEs strategy version
// duração de 1 hora, duração de 4 horas, em cerca de 12h,
// cerca de 12h, evolução 9h, por 24 h
RULENAME="duration_rafa5",
EXTRACTION="(dura[cç] [aã]o|cerca|evolu[cç] [aã]o|por)\s?
(de)?\s?(\d\d?\d?)\s?(horas?|hrs?|hs|h)",
NORM_VALUE="PTgroup(3)H"
```

```
// Example f3
// N-gram strategy version
// duração de 1 hora, duração de 4 horas, em cerca de 12h,
// cerca de 12h, evolução 9h, por 24 h, evol19h
// The "evol" abbreviation came from the N-gram strategy
RULENAME="duration_rafa5",
EXTRACTION="(dura[cç] [aã]o|cerca|evolu[cç] [aã]o|evol|por)\s?"
```

```
(de)?\s?(\\d\\d?\\d?)\s?(horas?|hrs?|hs|h)",
NORM_VALUE="PTgroup(3)H"
```

```
////////////////////////////////////
```

```
// Example g1
// Correct TEs strategy version
// 1 x dia, 1 x por dia, 1 x à noite, 1 x/d, 1 x/dia,
// 1 xdia, 1x / dia, 1x /dia, 1x a noite, 1x antes do café,
// 1x dia, 1x à noite, 1xd, 2 x / dia, 2 x dia, 2 x/d, 2x / dia,
// 2x/dia, 3 x/dia, 3x / dia, 3x / semana, 3x/dia
RULENAME="set_rafa1",
EXTRACTION="(\\d\\d?\\d?)\s?x\s?(/)?\s?%reSetUnitsRafa",
NORM_VALUE="P1%normSetUnitsRafa(group(3))",NORM_FREQ="group(1)X"
```

```
// Content of the reSetUnitsRafa file
```

```
dia
```

```
por dia
```

```
à noite
```

```
antes do café
```

```
após almoço
```

```
após jantar
```

```
semana
```

```
mês
```

```
// Content of the normSetUnitsRafa file
```

```
"dia","D"
```

```
"por dia","D"
```

```
"à noite","D"
```

```
"antes do café","D"
```

```
"após almoço","D"
```

```
"após jantar","D"
```

```
"semana","W"
```

```
"mês","M"
```

```
// Example g2
```

```
// Noisy TEs strategy version
```

```
// 1 x dia, 1 x por dia, 1 x à noite, 1 x/d, 1 x/dia,
```

```
// 1 xdia, 1x / dia, 1x /dia, 1x a noite, 1x antes do café,
```

```

// 1x dia, 1x à noite, 1xd, 2 x / dia, 2 x dia, 2 x/d, 2x / dia,
// 2x/dia, 3 x/dia, 3x / dia, 3x / semana, 3x/dia
RULENAME="set_rafa1",
EXTRACTION="(\d\d?\d?)\s?x\s?(/)?\s?%reSetUnitsRafa",
NORM_VALUE="P1%normSetUnitsRafa(group(3))",NORM_FREQ="group(1)X"

// Content of the reSetUnitsRafa file
d
dia
por\s?dia
[aà]\s?noite
antes\s?do\s?caf[eé]
ap[óo]\s?s?almo[çc]o
ap[óo]\s?s?jantar?
semana
m[eê]s

// Content of the normSetUnitsRafa file
"d","D"
"dia","D"
"por\s?dia","D"
"[aà]\s?noite","D"
"antes\s?do\s?caf[eé]","D"
"ap[óo]\s?s?almo[çc]o","D"
"ap[óo]\s?s?jantar?","D"
"semana","W"
"m[eê]s","M"

// Example g3
// N-gram strategy version
// 1 x dia, 1 x por dia, 1 x à noite, 1 x/d, 1 x/dia,
// 1 xdia, 1x / dia, 1x /dia, 1x a noite, 1x antes do café,
// 1x dia, 1x à noite, 1xd, 2 x / dia, 2 x dia, 2 x/d, 2x / dia,
// 2x/dia, 3 x/dia, 3x / dia, 3x / semana, 3x/dia
RULENAME="set_rafa1",
EXTRACTION="(\d\d?\d?)\s?x\s?(/)?\s?%reSetUnitsRafa",
NORM_VALUE="P1%normSetUnitsRafa(group(3))",NORM_FREQ="group(1)X"

// Content of the reSetUnitsRafa file

```

```

d
dia
por\s?dia
[aâ]\s?noite
antes\s?do\s?caf[eé]
ap[óó]\s?almo[çc]o
ap[óó]\s?jantar?
// Rafa add ngram
ap[óó]\s?jan
semana
// Rafa add ngram
sema
// Rafa add ngram
sem
m[eê]s

// Content of the normSetUnitsRafa file
"d","D"
"dia","D"
"por\s?dia","D"
"[aâ]\s?noite","D"
"antes\s?do\s?caf[eé]","D"
"ap[óó]\s?almo[çc]o","D"
"ap[óó]\s?jantar?","D"
// Rafa add ngram
"ap[óó]\s?jan","D"
"semana","W"
// Rafa add ngram
"sema","W"
// Rafa add ngram
"sem","W"
"m[eê]s","M"

```

```

////////////////////////////////////

```

```

// Example h1
// Correct TEs strategy version
// cp/dia, cp/manhã, cp/noite, cp / dia, cp a noite, cp a tarde,
// cp manhã, cp na hora do café da manhã, cp noite, cp pela manhã

```

```

// cp todos dias, cp vo no almoço, cp vo pela manhã, cp vo à noite,
// cp à noite,
// mcg / dia, mcg /dia, mg / d, mg / dia, mg /d, mg /dia, mg /noite
// mg a noite, mg dia, mg pela manhã, mg vo /dia,
// mg xdia, mg/d, mg/dia, pela noite, mg no almoço
// antes de dormir, antes do café, apos almoço, após o jantar
RULENAME="set_rafa4",
EXTRACTION="(ucg|cp|mcg|mg|vo|pela|antes|ap[oó]s|por)\s?
(de|do|vo|no|na|x|o|a)?\s?(/)?\s?%reSetUnitsRafa2",
NORM_VALUE="P1D"

// Content of the reSetUnitsRafa2 file
dias?
manhã
noite
à noite
a tarde
na hora do café da manhã
pela manhã
todos (os)? dias
vo no almoço
vo pela manhã
vo à noite
almoço
jantar
café
café da manhã
levantar
dormir

// Example h2
// Noisy TEs strategy version
// cp/dia, cp/manhã, cp/noite, cp / dia, cp a noite, cp a tarde,
// cp manhã, cp na hora do café da manhã, cp noite, cp pela manha,
// cp pela manhã
// cp todos dias, cp vo no almoço, cp vo pela manhã,
// cp vo à noite, cp à noite, cappelamanha
// mcg / dia, mcg /dia, mg / d, mg / dia, mg /d, mg /dia, mg /noite
// mg a noite, mg de manha, mg dia, mg pela manhã, mg vo /dia,

```

```

// mg xdia, mg/d, mg/dia, pela noite, mg no almoço
// antes de dormir, antes do café, apos almoço, após o jantar
RULENAME="set_rafa4",
EXTRACTION="(ucg|cp|mcp|mg|vo|pela|antes|ap[oó]s|por)\s?
(de|do|vo|no|na|x|o|a)?\s?(/)?\s?%reSetUnitsRafa2",NORM_VALUE="P1D"

// Content of the reSetUnitsRafa2 file
d
dias?
manhã
manha
noite
[ãa]\s?noite
a\s?tarde
na\s?hora\s?do\s?caf[eé]\s?da\s?manh[ãa]
pela\s?manh[aã]
todos\s?(os)?dias?
vo\s?no\s?almo[çc]o
vo\s?pela\s?manh[ãa]
vo\s?[ãa]\s?noite
almo[çc]o
jantar?
caf[eé]
caf[eé]\s?da\s?manh[ãa]
levantar
dormir

// Example h3
// N-gram strategy version
// cp/dia, cp/manhã, cp/noite, cp / dia, cp a noite, cp a tarde,
//cp manhã, cp na hora do café da manhã, cp noite, cp pela manha,
// cp pela manhã
// cp todos dias, cp vo no almoço, cp vo pela manhã, cp vo à noite,
// cp à noite, cappelamanha
// mcp / dia, mcp /dia, mg / d, mg / dia, mg /d, mg /dia, mg /noite
// mg a noite, mg de manha, mg dia, mg pela manhã, mg vo /dia,
// mg xdia, mg/d, mg/dia, pela noite, mg no almoço
// antes de dormir, antes do café, apos almoço, após o jantar
RULENAME="set_rafa4",

```

```

EXTRACTION="(ucg|cp|mcg|mg|vo|pela|antes|ap[oó]s|por)\s?
(de|do|vo|no|na|x|o|a)?\s?(/)?\s?%reSetUnitsRafa2",
NORM_VALUE="P1D"

// Content of the reSetUnitsRafa2 file
d
dias?
man(n)?h[ãa]
man(n)?ha
// Rafa altered - ngram - found "noide" in the training set
noi.e
// Rafa altered - ngram - found "noide" in the training set
[ãa]\s?noi.e
a\s?tarde
na\s?hora\s?do\s?caf[eé]\s?da\s?man(n)?h[ãa]
pela\s?man(n)?h[aã]
todos\s?(os)?dias?
vo\s?no\s?almo[çc]o
vo\s?pela\s?man(n)?h[ãa]
// Rafa altered - ngram - found "noide" in the training set
vo\s?[ãa]\s?noi.e
almo[çc]o
jantar?
// Rafa added - ngram - found "jan" in the training set (jantar - diner)
jan
caf[eé]
caf[eé]\s?da\s?manh[ãa]
levantar
dormir

////////////////////////////////////

// Example i1
// Correct TEs strategy version
// às 8:00, às 16:00 e às 20:00, às 8:00 às 16:00 e às 20:00
RULENAME="set_rafa64",
EXTRACTION="(às)\s?(%reTimeHour:%reTimeMinute)\s?(,)?\s?(às)?\s?
(%reTimeHour:%reTimeMinute)\s?(e)?\s?(às)?\s?(%reTimeHour:%reTimeMinute)",
NORM_VALUE="P1D",

```

```

NORM_FREQ="3X"

// Example i2
// Noisy TEs and N-gram strategy versions
// às 8:00, às 16:00 e às 20:00, às8:00as16:00eàs20:00
RULENAME="set_rafa64",
EXTRACTION="(às|ás|as)\s?(%reTimeHour:%reTimeMinute)\s?(,)?\s?
(às|ás|as)?\s?(%reTimeHour:%reTimeMinute)\s?(e)?\s?(às|ás|as)?\s?
(%reTimeHour:%reTimeMinute)",
NORM_VALUE="P1D",
NORM_FREQ="3X"

```

The amount of true positive (TP) found in the extraction step that were evaluated concerning the normalization step is presented in table 10.

Table 10 – Quantity of TP TEs extracted by HeidelTime which were used to evaluate the normalization step

Experiment	Extraction
	TP
Baseline	459
Correct TEs	671
Noisy TEs	820
N-gram TEs	824

The sample standard deviation of extraction and normalization steps made by HeidelTime considering true positives (TP) and false positive (FP) values in each experiment are presented in table 11. The calculation was made using the same 6 test files with 20 clinical texts each file for each experiment.

Table 11 – Sample standard deviation of extraction and normalization steps made by HeidelTime in each experiment

Experiment	Extraction		Normalization		Both
	TP	FP	TP	FP	FN
Baseline	8,60	5,28	9,21	5,54	12,56
Correct TEs	10,74	2,48	10,26	4,55	13,34
Noisy TEs	12,44	5,04	12,85	4,05	2,74
N-gram TEs	13,88	7,86	12,21	5,28	2,94

In table 11, the column called "Both" indicates the false negative (FN) value for both steps, extraction and normalization. This decision was made having in mind that, if a TE is FN in the extraction step, it is also a FN in the normalization step. We assumed

a practical interpretation of the fact, meaning if the TE was not extracted by the tagger, it affects the system that would use it, as systems only understand normalized TEs. This information is important, as the same FN values were used in the extraction and normalization steps to calculate the recall and consequently the F1-Score. It is important to have in mind that, there is no FN in the normalization step, as it depends totally on the extraction step.

Intuitively, one can think that, as an incremental approach was chosen, the last experiment should present the best results. However, the results showed that this assumption did not hold true. Table 12 and table 13 show the results of all experiments of the extraction and normalization steps respectively.

Table 12 – Extraction experiments results

Experiment	Extraction		
	Precision	Recall	F1-Score
1 - Baseline	0.8500	0.5550	0.6715
2 - Correct TEs	0.8971	0.7695	0.8284
3 - Noisy TEs	0.8659	0.9414	0.9021
4 - N-gram TEs	0.8400	0.9471	0.8903

Table 13 – Normalization experiments results

Experiment	Normalization		
	Precision	Recall	F1-Score
1 - Baseline	0.7352	0.5027	0.5971
2 - Correct TEs	0.8815	0.7475	0.8090
3 - Noisy TEs	0.9127	0.9366	0.9245
4 - N-gram TEs	0.9028	0.9424	0.9221

Tables 12 and 13 show how metrics changed its values in each experiment for both step. Considering table 12, the Precision value of the experiments two and three are close. The same behavior is observed with the Recall and F1-Score results of the experiments three and four. Precision, Recall and F1-Score of experiments three and four presented in table 13 also have close values. To be sure that there was a significant increase of quality between the best result of each metric (in bold) and the other results of the same metric, statistical tests were made with a confidence level of 0.95 and a significance level of 0.05.

As already mentioned, all experiments were made over the same conditions (paired samples). The Shapiro-Wilk normality test showed data distribution was not normal. This test was chosen because the number of samples was less than 50 (RAZALI; WAH et al., 2011). The tests were made per test file (6 in total), see details in table 5. Aware of this dataset feature, two non-parametrics tests were run, Wilcoxon Test and Friedman Test

with Dunn-Bonferroni post-hoc Test² (DUNN, 1964). The results of the tests are shown in tables 14 to 19.

Table 14 – Significance of Wilcoxon Test in **extraction**

Statistical difference in extraction (Wilcoxon Test) evaluating F1-Score of each test file (there is difference if significance <0.05 - in bold)				
Experiment	1 - Base.	2 - Corr.	3 - Nois.	4 - N-gr.
1 - Base.	#	0.028	0.028	0.028
2 - Corr.	#	#	0.028	0.046
3 - Nois.	#	#	#	0.046
4 - N-gr.	#	#	#	#

Table 14 shows the significance values of the Wilcoxon Test. All bold values show that there is a significant statistical difference between both tests. The results show experiments two, three and four are statistically different from the first experiment. It means experiments two, three and four have statistically better extraction quality than the first experiment. The table also shows that experiments three and four are better than the second experiment. Finally, the Wilcoxon Test for the extraction step shows the experiment four has a better extraction quality than the experiment three. However, from a significance point of view, to consider one experiment better than the other, the significance value of the Wilcoxon Test must be lower than 0.05. Having said that, the 0.046 values in the table show that there was a tiny statistical difference between these experiments. To check the statistical difference of the experiment in a more accurate way, the Friedman Test with the Dunn-Bonferroni post-hoc Test was applied for all experiments.

Table 15 – Significance of Friedman and Dunn-Bonferroni tests in **extraction** for all experiments

Statistical difference in extraction (Friedman + Dunn-Bonferroni tests) evaluating F1-Score of each test file (there is difference if significance <0.05 - in bold)				
Experiment	1 - Base.	2 - Corr.	3 - Nois.	4 - N-gr.
1 - Base.	#	0.705	0.001	0.044
2 - Corr.	#	#	0.152	1.000
3 - Nois.	#	#	#	1.000
4 - N-gr.	#	#	#	#

Results showed by the Friedman + Dunn-Bonferroni tests (table 15) showed there is a statistical difference only between the experiment one and the experiments three and four. However, the significance between the experiments one and four was also a tiny

² <http://www-01.ibm.com/support/docview.wss?uid=swg21508972>

difference. To investigate this difference deeper, a test only with experiments two, three and four was done.

Table 16 – Significance of Friedman and Dunn-Bonferroni tests in **extraction** for 3 experiments

Statistical difference in extraction (Friedman + Dunn-Bonferroni tests) evaluating F1-Score of each test file (there is difference if significance <0.05 - in bold)			
Experiment	2 - Corr.	3 - Nois.	4 - N-gr.
2 - Corr.	#	0.012	0.447
3 - Nois.	#	#	0.447
4 - N-gr.	#	#	#

The Friedman + Dunn-Bonferroni tests (table 16) showed that the only statistical difference in the extraction step existed only between the experiment two (2 - Corr.) and experiment three (3 - Nois.). This is the result taken as the final result of the extraction step, meaning there is a statistical difference only between the experiment number two (2 - Correct TEs) and experiment number three (3 - Noisy TEs). It means there is no statistical difference between the experiment number three (3 - Noisy TEs) and experiment number four (4 - N-gram TEs). So, the best result of the extraction step considering the F1-Score is 0.9021, which is also the highest result among all experiments (table 12).

Table 17 – Significance of Wilcoxon Test in **normalization**

Statistical difference in normalization (Wilcoxon Test) evaluating F1-Score of each test file (there is difference if significance <0.05 - in bold)				
Experiment	1 - Base.	2 - Corr.	3 - Nois.	4 - N-gr.
1 - Base.	#	0.028	0.028	0.028
2 - Corr.	#	#	0.028	0.028
3 - Nois.	#	#	#	0.600
4 - N-gr.	#	#	#	#

Table 17 shows the significance values of the Wilcoxon Test. All bold values show that there is a significant statistical difference between all tests, except between experiments three and four. Results show experiments two, three and four are statistically different from the first experiment. It means experiments two, three and four have statistically better normalization quality than the first experiment. The table also shows that experiments three and four are better than the second experiment. Finally, the Wilcoxon Test for the normalization step shows experiments three and four have the same normalization quality. To check the statistical difference of the experiment in a more accurate way, the Friedman Test with the Dunn-Bonferroni post-hoc Test was applied for all experiments. Results are presented in table 18.

Table 18 – Significance of Friedman and Dunn-Bonferroni tests in **normalization** for all experiments

Statistical difference in normalization (Friedman + Dunn-Bonferroni tests) evaluating F1-Score of each test file (there is difference if significance <0.05 - in bold)				
Experiment	1 - Base.	2 - Corr.	3 - Nois.	4 - N-gr.
1 - Base.	#	1.000	0.002	0.010
2 - Corr.	#	#	0.152	0.442
3 - Nois.	#	#	#	1.000
4 - N-gr.	#	#	#	#

Table 18 shows there is a statistical difference between experiments three (3 - Nois.) and four (4 - N-gr.) in relation to the experiment one (1 - Base.). Table 19 shows results of Friedman and Dunn-Bonferroni tests excluding the first experiment.

Table 19 – Significance of Friedman and Dunn-Bonferroni tests in **normalization** for 3 experiments

Statistical difference in normalization (Friedman + Dunn-Bonferroni tests) evaluating F1-Score of each test file (there is difference if significance <0.05 - in bold)			
Experiment	2 - Corr.	3 - Nois.	4 - N-gr.
2 - Corr.	#	0.012	0.063
3 - Nois.	#	#	1.000
4 - N-gr.	#	#	#

As a final consideration of the statistical test in the normalization step, table 19 points out the only statistical difference among all experiments is between experiments two (2 - Correct TEs) and three (3 - Noisy TEs). Thus, the best F1-Score pointed out for the normalization step, 0.9245, is validated with the statistical tests. We reached this conclusion because Friedman and Dunn-Bonferroni tests result in table 19 does not present a significant difference between experiments three (3 - Noisy TEs) and four (4 - N-gram TEs).

In summary, experiment three (3 - Noisy TEs) presented the best results in this research, considering extraction and normalization steps of the Temporal Tagging Task. The best F1-Score of the extraction step was 0.9021. The best F1-Score of the normalization step was 0.9245. The next chapter presents the discussion of this study.

8 Discussion

This chapter discusses the results obtained in the application of the proposed method. It includes mostly, situations and comments about the results obtained with the comparison between TEs of the test set (ground truth) and HeidelTime setup with the N-gram rules version. As original texts are in Portuguese, Brazilian health professionals tried to translate each snippet to English, as our work is written in English. However, there are some cases that even these professionals could not understand these non-standard abbreviations. Because of the non-clear understanding of these non-standard abbreviations, someone may disagree about some non-standard abbreviations expansion. Text within "[...]" are necessary explanations related to the abbreviations found in some clinical reports. The corpus used was de-identified, so names like Lucas, Florence and others appears so many times. These are not the true names of patients.

I. Texts below are examples of True Positives in both steps, extraction and normalization. When the TE is only bold, it indicates HeidelTime tagged it identically with the test set. The part of the TE which is underlined was tagged only by HeidelTime. When there is an underlined part, the bold part indicates the TE tagged in the test set.

a) "paciente de **75 anos**, cardiopata isquêmico (iam inf **mar/02** - actp + stent em cx; sca s/ st em dez/03) hipertenso prévio, dislipidêmico, portador de insuficiência renal crônica e gota . vem transferido ao cti / xxxx no dia 28/02/2004 por dispnéia progressiva e angina . iniciado tratamento clínico para sca e edema agudo de pulmão, obtendo-se boa resposta . apresentou elevação discreta de ck-mb (32) , sem alterações compatíveis com isquemia recente ao ecg . submetido, em 02/03/2004, a avaliação angiográfica, que evidenciou restenose intra-stent e proximal ao stent em dae e lesão em 1a diagonal . realizada angioplastia de artéria diagonal e actp + stent em da, com bom resultado angiografico . o paciente teve boa evolucao após o procedimento, recebe alta em boas condições clínicas, estável . orientação de manter medicações acima relacionadas, e utilizar ticlopidina 250 mg **12/12 h** por 30 dias . retorna ao ambulatório da cardio (previamente já em acompanhamento) ."

1) Snippet: "paciente de **75 anos**, cardiopata isquêmico (iam inf [Infarto agudo do miocárdio inferior] **mar/02** - actp + stent [angioplastia coronária transluminal percutânea + stent (é uma endoprótese expansível circunflexa)] em cx; sca s/ st [circunflexa; síndrome coronariana aguda sem st (st é um segmento do eletrocardiograma. É a porção do ECG entre o final do complexo QRS e o começo da onda T)] em dez/03) hipertenso prévio"

- 2) Free translation: "patient of **75 years**, ischemic heart disease (iam inf [inferior / acute myocardial infarction] **mar/02** - actp + stent [percutaneous transluminal coronary angioplasty + stent (is a circumflex expandable endoprosthesis)] in cx; sca s/ st [circumflex, acute coronary syndrome without st (st is a segment of the electrocardiogram. Is the portion of the ECG between the end of the QRS complex and the beginning of the T wave] in 10/03) previous hypertension"
 - 3) Test set: 75 anos | DURATION | P75Y
HeidelTime: 75 anos | DURATION | P75Y
 - 4) Test set: mar/02 | DATE | 2002-03
HeidelTime: mar/02 | DATE | 2002-03
 - 5) Snippet: "orientação de manter medicações acima relacionadas, e utilizar ticlopidina 250 mg 12/12 h por 30 dias ."
 - 6) Free translation: "to maintain medications listed above, and to use ticlopidine 250 milligrams each 12/12 hours for 30 days."
 - 7) Test set: 12/12 h | SET | P1D | 2X
HeidelTime: mg 12/12 | SET | P1D | 2X
 - 8) Discussion: TEs **75 anos** and **mar/02** are examples of the correct extraction and normalization made by HeidelTime, where both steps are identical to the test set. TEs **12/12 h** and **mg 12/12** are also correct in the extraction and normalization steps, once the relaxed matching approach was used for the extraction step (see section 2.3 for more details). The TE **mg 12/12** shows that, the rule created in HeidelTime included the **mg** not to confused **12/12** meaning twice a day, instead of December 12 or December/2012.
- b) "# ic etiologia isquemica # iam em 27/08/2010 # pot atc c stent # pot rvm 2011 mie - da / sf - diag em uso de carvedilol **50 / dia** + rosuva 10 + valsartan 320 + furo 40 + apresolina 100 / dia + espiro 25 + aas 100 + monoc 20 / dia + mtf 850 3x + levotirox 50 + indapamida 1, 5 s # estável, nega precordialgia, refere dispneia aos grandes esforços . sem edema de mmii, mantem uso regular de medicações . refere somente perda da potencia sexual trouxe nova cintilo (06/10/14) : hipoperfusão persistente (fibrose) de grande extensão, no segmento meio apical da parede anterior e ápice de ve, em grau discreto na parede anteroseptal fe repouso 49% o # pa 110x70 fc 79 sat 98% corado, hidratado, euvolemico, eupneico bcrnf ss 2t mv+ sim bilart sra abd sp mmii s edema a # ic isquemica p # mantenho medicações retorno em 6 meses com ecott oriento diminuir gradativamente o monocordil, caso suporte sem dispneia, ver possibilidade do uso de cialis / viagra ."
- 1) Snippet: "pot rvm [Pós operatório tardio de Revascularização no miocárdio] 2011 mie - da / sf - diag [Mamária interna esquerda – descendente anterior

/ sf- diagonal] em uso de carvedilol **50 / dia** + rosuva [Rosuvastatina] 10 + valsartan 320"

- 2) Free translation: "pot rvm [Late postoperative myocardial revascularization] 2011 mie - da / sf - diag [left internal mammary - anterior descending / sf - diagonal] using carvedilol **50 / day** + rosuva [Rosuvastatin] 10 + valsartan 320"
 - 3) Test set: 50 / dia | SET | P1D
HeidelTime: 50 / dia | SET | P1D
 - 4) Discussion: In this case, Heideltime made the extraction and normalization correctly. The rule was created to include a number and the slash in order to avoid mistakes with the word "dia" (day) alone.
- c) "primeira consulta irc - nefroesclerose hipertensiva? com gesf secundária? tgf estimada em 20 hematuria - com história de tabagismo lucas, 53 anos - paciente encaminhada por piora de função renal - queixa de astenia - portadora de has com diagnóstico **há cerca de 4 anos**, com controle regular - hipotireoidismo com dx há um ano . - nega cirurgias prévias - itus prévias uso de: losartan 50mg 12/12h, furosemida 40mg, levotiroxina 100mcg, omeprazol 20mg . - ex tabagista 40 anos maço, parou **há 5 anos** - sedentária - nega história de doença renal na família trouxe exames laboratoriais: - hb 12 - creatinina 2, 7 (tfg estimada 20ml/min) - hematuria - us - rins 9, 2 e 9, 4, com cistos simples (3 no total) , bexiga normal - 2012 - proteinuria de 500mg ao exame: corada, hidratada, afebril pa 140/90 mv positivo, simetrico abdomen sp"
- 1) Snippet: "queixa de astenia - portadora de has com diagnóstico **há cerca de 4 anos**, com controle regular - hipotireoidismo com dx há um ano ."
 - 2) Free translation: "complaint of asthenia - patient with systemic arterial hypertension (has) diagnosed for **about 4 years**, with regular control - hypothyroidism with diagnosis (dx) one year ago."
 - 3) Test set: há cerca de 4 anos | DATE | 2008
HeidelTime: há cerca de 4 anos | DATE | TPZ-0004
 - 4) Snippet: "ex tabagista 40 anos maço, parou **há 5 anos** - sedentária"
 - 5) Free translation: "former smoker 40 years pack, stopped **5 years ago** - sedentary"
 - 6) Test set: há 5 anos | DATE | 2007
HeidelTime: há 5 anos | DATE | TPZ-0005
 - 7) Discussion: In both TEs **há cerca de 4 anos** and **há 5 anos** the extraction and normalization are correct. The normalization in both cases are correct, even

though, they are different, because they are equivalent. TPZ stands for (Time Point Zero) (STRÖTGEN, 2015), the signal "-" indicate a past time (it also can be "+", which indicates future time). TPZ-0004 indicates the event "has" (hipertensão arterial sistêmica - systemic arterial hypertension) was identified 4 years ago. The same logic applies to the normalization TPZ-005, which is associated with the event "ex tabagista" (former smoker). The rule with the TPZ normalization already existed in HeideTime version downloaded from GitHub.

d) "lucas 28 anos gnc, perda renal pós-gravidez, bx realizada > não sabe laudo has drc, em trs - hd junho 2015 nega dm, dlp, tiróide, cardiopatias, litíase, outros . . . g 2 a 1 c 1 . nega outras cirurgias, diurese de bom volume . em uso de renalvit, epo, calciolit . internações > últimas por infecção permcath . trocado em nov huc e retirado em dez por febre . hoje com dl em jid, sem queixas **há mais de 1 m, ês** . peso seco 66, alt 160 . ao exame : pa 150/90, mucosas hipocoradas, demais exame físico normal . liberada para transplante, esclareço riscos e benefícios, abo o, sorologias cmv e ebv reagente, antihbs reagente, toxo nr, ecocardiograma normal, doppler pelvico normal, ecografia abdome total > rins nefropatia, demais sp, ecg sinusal, fc 113, normal, te ausencia alterações isquêmicas, imuno : rever mm e pra na reunião . "

- 1) Snippet: "hoje com dl em jid [nenhum dos profissional da saúde envolvidos nesta pesquisa soube o significado dessas abreviações **dl em jid**], sem queixas **há mais de 1 m, ês** . peso seco 66, alt 160 . "
- 2) Free translation: "today with dl in jid [none of the health professionals involved in this research knew the meaning of these abbreviations **dl in jid**], no complaints for more than 1 month [m, ês - written in a noisy way]. dry weight 66, height 160."
- 3) Test set: há mais de 1 m, ês | DURATION | P1M
HeideTime: há mais de 1 m | DATE | TPZ-0000-01
- 4) Discussion: TEs **há mais de 1 m, ês** and **há mais de 1 m** are extracted and normalized correctly. It is an example of noisy TEs, because **há mais de 1 m, ês** is wrongly written, the correct form would be **há mais de 1 mês**. However, as the letter "m" is the abbreviation of "mês" (month) in clinical reports, the HeideTime rule extracted it correctly. The confusion between types DATE and DURATION is also reported by Hagège, Baptista and Mamede (2008) and Viani et al. (2018).

e) "lucas, 62 anos comorbidades: # mcp dilatada / chagas # has # dislipidemia # marca-passo há 5 anos # dac familiar chv: ex-tabagista de 38 maços-ano hma:

paciente com icc classe funcional ii retorna com resultado do cat solicitado na última consulta; relata melhora do edema desde a última consulta/ cat: do dia 29/07/2014: protocolo pré - transplante cardíaco realizado com sucesso; rvp wood positiva: candidato atransplante; medicamentos: # enalapril 10 mg; 1/2 cp de 12/12 horas # apresolina 25 mg 1 cp de 8/8 horas # monocordl 20 mg 1 cp **as 8 e outro as 14 horas** # espironolactona 25 mg 1 **cp ao dia** # furosemida 40 mg 1 cp as 8 e outro as 14 horas # carvedilol 25 mg 1 cp de 12/12 horas # sinvastatina 20 mg 1 **cp a noite** conduta: solicitado internamento com monitorização para realização de transplante ;"

- 1) Snippet: "# monocordl 20 mg 1 cp **as 8 e outro as 14 horas** # espironolactona 25 mg 1 **cp ao dia** # furosemida 40 mg"
 - 2) Free translation: # monocordil 20 mg 1 capsule **at 8 and another at 14 hours** # spironolactone 25 mg 1 **capsule daily** # furosemide 40 mg
 - 3) Test set: as 8 e outro as 14 horas | SET | P1D | 2x
HeidelTime: as 8 e outro as 14 horas | SET | P1D | 2x
 - 4) Test set: cp ao dia | SET | P1D
HeidelTime: ao dia | SET | P1D
 - 5) Snippet: "# sinvastatina 20 mg 1 **cp a noite** conduta: solicitado internamento"
 - 6) Free translation: # simvastatin 20 mg 1 **capsule per night** conduct: requested admission
 - 7) Test set: a noite | SET | P1D
HeidelTime: cp a noite | SET | P1D
 - 8) Discussion: Three different TEs where used as examples in this clinical text. The TE **as 8 e outro as 14 horas** extraction and normalization made by HeidelTime was exactly the same as found in the test set. However, the TE **cp ao dia** and **ao dia** (test set and HeidelTime, respectively) is almost the same, being considered a correct extraction because of the relaxed matching approach. It shows a case where the size of the TE in the test set was longer than the TE extracted by the tagger. The TE **a noite** and **cp a noite** (test set and HeidelTime, respectively) presents the opposite situation, where the TE in the test set is shorter than the one tagged by HeidelTime. In all cases, the normalization was identical to the test set.
- f) "paciente florence nightingale 80 anos, vem transferido de curitiba (sem nota de transferência) , onde esteve internada segundo familiar por dor torácica e dispnéia . foi transferida do hospital de origem para a hemodinâmica, aos cuidados do prof . lucas a fim de realizar cateterismo cardíaco em 03/03/2006 - apresentou após procedimento edema agudo de pulmão . realizou ecocardiograma em 06/03/2006 - ve

dilatado, fe 31%, insuficiência mitral moderada funcional, e acinesia de segmentos anterior, medio, apical e inferior . em 01/03/2006 realizou angioplastia com implante de stent na primeira diagonal . em 07/03/2006 apresentou diminuição de volume de diurese e sintomas de uremia associado a hipervolemia necessitando de suporte dialítico por cateter . após colocação de svd apresentou hematúria maciça com queda de hb de 12 -> 8 em 24 horas . realizado tc de abdome que excluiu hipótese de hematoma retroperitoneal . foi transferida à uti em 08/03/2006 para otimização de manejo e monitorização, tendo alta para enfermaria dois dias após . melhora progressiva de sintomas - persistindo em insuficiência cardíaca classe 4 . paciente é transferida **hoje** para hospital de origem para seguir aos cuidados de seu médico, devendo manter acompanhamento com nefrologista também . prescrição aas 100 mg/dia ticlopidina 250 mg 2x/dia por 30 dias digoxina 0, 125 mg/ 48 horas hidralazina 25 mg 3x/dia isossorbida 20 mg 3x/dia carbonato de cálcio 500 mg caj hemodiálise por cateter 2/4/6 lista de problemas # insuficiência cardíaca classe 4 # cardiopatia isquêmica # doença renal crônica agudizada por nefropatia relacionada ao contraste # hematúria a esclarecer"

- 1) Snippet: "paciente é transferida **hoje** para hospital de origem para seguir aos cuidados de seu médico,"
 - 2) Free translation: "patient is transferred **today** to hospital of origin to follow the care of his doctor,"
 - 3) Test set: hoje | DATE | XXXX-XX-XX
HeidelTime: hoje | DATE | PRESENT_REF
 - 4) Discussion: The TE **hoje** was normalized to **XXXX-XX-XX** in the test set, but was normalized to **PRESENT_REF** by HeidelTime. This rule already existed in the downloaded version of the tagger. It is considered correct, because the present date is in relation to the reference date found by HeidelTime, which is identified automatically when the scientific domain is setup. Even if the annotator did not find the date, the word "hoje" indicates the present date when the text was written.
- g) "paciente com angina há mais de 2 anos, cintilografia miocárdica demonstrando hipoperfusão ínfero lateral e antero basal e cateterismo cardíaco de 21/06/06 com lesão crítica de 95% na coronária circunflexa distal e lesão de 70% na coronária direita média . internado para observação após angioplastia com implante de prótese intraluminal na lesão descrita na circunflexa, com sucesso angiográfico e sem intercorrências . recebe alta com plano de reintervenção percutânea **em uma semana** para tratamento da lesão na coronária direita ."

- 1) Snippet: "recebe alta com plano de reintervenção percutânea **em uma semana** para tratamento da lesão na coronária direita ."
 - 2) Free translation: "is discharged with a percutaneous reintervention plan **within a week** to treat the right coronary lesion."
 - 3) Test set: em uma semana | DATE | 2006-06-28
HeidelTime: em uma semana | DATE | FUTURE_REF
 - 4) Discussion: The normalization is correct, even being less accurate, because the TE "**em uma semana**" indicates a future date.
- h) "paciente hipertenso, cardiopata isquêmico, história de iam em parede inferior **há 14 dias** atrás com actp primária e implante de stent em cd, interna para realização de actp + stent em lesão grave em acx . procedimento realizado sem intercorrências, com sucesso angiográfico (fluxo timi 3) e sem lesão residual . tem alta hospitalar em bom estado geral, com indicação de seguir acompanhamento com médico assistente . medicações na alta captopril 50 mg 3x/d aas 200 mg/d ticlopidina 250 mg 12/12h por 3 meses hidroclorotiazida 25 mg/d"
- 1) Snippet: "paciente hipertenso, cardiopata isquêmico, história de iam (Infarto agudo do miocárdio) em parede inferior **há 14 dias** atrás com actp [angioplastia coronária transluminal percutânea] primária e implante de stent [(é uma endoprótese expansível circunflexa)] em cd [coronária direita]"
 - 2) Free translation: "hypertensive patient, ischemic heart disease, history of iam (acute myocardial infarction) in the lower wall **14 days** ago with actp primary [percutaneous transluminal coronary angioplasty] and stent implantation [is a circumflex expandable endoprosthesis] in cd [right coronary] "
 - 3) Test set: há 14 dias | DATE | XXXX-XX-XX
HeidelTime: há 14 dias | DATE | TPZ-0000-00-14
 - 4) Discussion: This case is also an example of the use of the TPZ rule implemented in the downloaded version of HeidelTime, which is more precise than the normalization found in the test set.
- i) "l . o . , **58 a** . dm diagnosticada há 5 anos, em uso de metformina 850mg 3x / dia e glicazida 2x / dia (sem tomar há 2 meses) has há 14 anos, em uso de enalapril, losartana (sem tomar há 2 dias) , hidroclorotiazida e alondipino . em uso de doxazosina 4mg, para hpb . refere cx na bexiga devido a tumor (sic) em 2013 . sem queixas associadas, encaminhado para acompanhamento de dm2 . nega alterações de urina ou de visão . refere dor em calcanhar direito, sem presença de lesões . refere dieta adequada, sem consumo excessivo de doces ou gorduras . sedentário . refere emagrecimento de 17 kgs em 2, 5 anos . nega etilismo . ex-tabagista, 23

maços / ano, parou há 20 anos . refere alergia a aas . hf : tias com dm . mãe has . lab (07/05/15) : hemograma sem alterações cr 0, 9 glicose 241 acido urico 7, 3 transaminase piruvica 42 hdl 30 k 4, 2 transaminase oxalacetica 26 ct 244 tsh 1, 55 hb1ac : 11% microalbuminuria 24h : 177mg (2500ml) tg : 657mg/dl ex fisico : pa: 180/130 fc : 80 fr: 18 sato2 : 93 peso : 101, 70kg altura : 1, 63m imc : 38, 28 ca : 121cm precordio : bcrnfss pulmonar : mv +, sem ra abdome : globoso, flacido, rha+, indolor a palpação mmii : sem edema, sem lesões em pés cd : prescrevo insulina nph 20u a noite, metformina 850mg 3x / dia e glibenclamida 5mg 2x / dia . enalapril 10mg 2x / dia, anlodipino 5mg 2x / dia, hidroclorotiazida 25mg 1x / dia, sinvastatina 20mg 1x / dia solicito exames laboratoriais"

- 1) Snippet: "l. o . [abreviação do nome do paciente], **58 a** . dm [diabetes mellitus] diagnosticada há 5 anos"
 - 2) Free translation: "l. o . [abbreviated patient's name], **58 years** . dm [diabetes mellitus] diagnosed 5 years ago"
 - 3) Test set: 58 a | DURATION | P58Y
HeidelTime: 58 a | DURATION | P58Y
 - 4) Discussion: The TE **58 a** was correctly extracted (and correctly normalized), because a rule which extracts this common abbreviation in clinical report was implemented in HeidelTime.
- j) "# has # actp + stent da media pct 64 anos, com angina classe 2, foi submetido a angioplastioa com implante de stent na arteria coronaria descendente anterior . procedimento ocorreu sem intercorrencias com sucesso angiografico . o pct recebe alta em boas condições clinicas para acomapanhamento ambulatorial apos 24 h de observação . # # recomendações pós-alta # # retorno a emergencia em caso de qualquer intercorrenci -uso de ticlopidina **250mg 2x por 1 mes** rretorno ao ambulatorio de cardiologia conforme agendado manter medicações em uso"
- 1) Snippet: "retorno a emergencia em caso de qualquer intercorrenci -uso de ticlopidina **250mg 2x por 1 mes** rretorno ao ambulatorio"
 - 2) Free translation: "return to emergency in case of any intercurrance - use of ticlopidine 250 **milligrams, twice daily for 1 month**, return to the outpatient clinic"
 - 3) Test set: mg 2x | SET | P1D | 2X
HeidelTime: 2x | SET | P1D | 2X
 - 4) Test set: por 1 mes | DURATION | P1M
HeidelTime: 1 mes | DURATION | P1M
 - 5) Discussion: The TE **2x** was considered a correct extraction and normalization, because in most of the cases it indicates twice a day. However, in some less

common cases, it could indicate twice a week or a month, for example. The TE **1 mes** was also correctly extracted and normalized.

II. The following examples are cases of False Positives, which were not TEs, but HeidelTime extracted them as TEs.

a) "paciente de 75 anos, cardiopata isquêmico (iam inf mar/02 - actp + stent em cx; sca s/ st em dez/03) hipertenso prévio, dislipidêmico, portador de insuficiência renal crônica e gota . vem transferido ao cti / xxxx no dia 28/02/2004 por dispnéia progressiva e angina . iniciado tratamento clínico para sca e edema agudo de pulmão, obtendo-se boa resposta . apresentou elevação discreta de ck-mb (32) , sem alterações compatíveis com isquemia recente ao ecg . submetido, em 02/03/2004, a avaliação angiográfica, que evidenciou restenose intra-stent e proximal ao stent em dae e lesão em 1a diagonal . realizada angioplastia de artéria diagonal e actp + stent em da, com bom resultado angiografico . o paciente teve boa evolucao após o procedimento, recebe alta em boas condições clínicas, estável . orientacao de manter medicações acima relacionadas, e utilizar ticlopidina 250 mg 12/12 h por 30 dias . retorna ao ambulatório da cardio (previamente já em acompanhamento) ."

1) Snippet: "que evidenciou restenose intra-stent e proximal ao stent em dae [Descendente anterior esquerda] e lesão em 1a diagonal."

2) Free translation: "which showed restenosis in-stent and proximal to the stent in dae [left anterior descending] and in 1st diagonal lesion."

3) HeidelTime: "type": "DATE", "value": "FUTURE_REF", "text": "em 1a"

4) Discussion: "em 1a" is an abbreviation of 1 year found in the training set and became a new rule in HeidelTime. The preposition "em" is also part of the TE, in order not to create a confusion between "1a", which is the same as 1st in English.

b) "paciente interna com iam anterior kiliip ii com supra st, realizado actp primaria na setima hora + implante de stent, com criterios para reperfusão porém com formação de onda q e evolução para quadro transitorio de baixo debito, congestão pulmonar, acfa, bre intermitente, todos tratados clinicamente . melhora sintomatica em 48h com tto clinico, já sem congestão, classe funcional ii / iii . presença de taquicardia constante e presença de b3, preditoras de ice pos iam . ecocardio pos iam mostrou fe 44% com acinesia total da parede anteroseptal, sem complicações mecanicas . iniciado betabloqueador na internação com boa tolerancia . alta com tto otimizado de icc, em classe funcional ii, b3 persistente, sem sinais de congestão ao exame fisico e rx . mortalidade calculada em 30 dias pelos criterios timi em 27%, mesmo com

todo tto oferecido, risco que foi explicado a familia e orientado o paciente para procurar atendimento se necessario ."

- 1) Snippet: "mortalidade calculada em 30 dias pelos criterios timi em 27%, mesmo com todo tto oferecido"
 - 2) Free translation: "mortality calculated by the timi criteria in 30 days by 27%, even with all treatment offered"
 - 3) HeidelTime: "type": "DATE", "value": "1927", "text": "em 27"
 - 4) Discussion: In this case, the rule created to extract a year typed with two digits, like "em 94", meaning the year 1994, was incorrectly extracted, as "em 27" was not a year, but it was a percentage.
- c) "homem, 69 anos, retorno com exames . hmp : dm há 10 anos, has há 10 anos e dislipidemia . faz uso de sinvastatina 30 mg, ferro 250 mg, diovan 160 mg, metformina 850 mg, aas 1 cp no almoço . cx de prótese de quadril em março de 2013 e cx para correção de fratura em calâneo há 10 anos . nega pé diabético, nega edema . hmf : pai hipertenso, tio com ca de estômago chv : ex tabagista 45 anos / maço (parou há 10 anos) rs : dificuldade para controlar esfíncter - não controla evacuação . exame físico : beg, lote, corado, hidratado pa : 139/75 mmhg fc : 75 bpm peso; 79 kg altura : 1, 63 m exames : creatinina : 1, 5 uréia : 66 ckd/epi : 47 hemoglobina : 10, 8 ht : 33, 4% leucócitos : 13050 glicemia : 90 calcio ionico : 1, 28 fosforo : 4, 3 potassio : 4, 8 sodio : 138 ácido úrico : 7, 1 conduta : pedidos exames - hba1c - glicemia - relação albumina / creatinina em amostra de urina isolada - ferritina - irst - pth - 25o h vitamina d - lipidograma"
- 1) Snippet: "pai hipertenso, tio com ca [câncer] de estômago chv [condição e hábitos de vida]: ex tabagista 45 anos / maço (parou há 10 anos) rs [revisão de sistemas]: dificuldade para controlar esfíncter"
 - 2) Free translation: "hypertensive father, uncle with ca [cancer] stomach chv [condition and habits of life]: ex smoker 45 years / pack (stopped 10 years ago) rs [revision of systems]: difficulty to control sphincter"
 - 3) HeidelTime: "type": "DURATION", "value": "P45Y", "text": "45 anos"
 - 4) Snippet: "pa [pressão arterial] : 139/75 mmhg [milímetros de mercúrio] fc [frequência cardíaca]: 75 bpm peso; 79 kg altura : 1, 63 m exames : creatinina : 1, 5"
 - 5) Free translation: "pa [blood pressure]: 139/75 mmhg [millimeters of mercury] fc [heart rate]: 75 bpm [beats per minute] weight; 79 kg height: 1, 63 m exams: creatinine: 1, 5 "
 - 6) HeidelTime: "type": "DURATION", "value": "PT63M", "text": "63 m"

- 7) Discussion: The rule extracted the text 45 anos, because taking it isolated, it can be understood as a TE of the type DURATION. However, considering the context, this type of text was not supposed to be annotated (see item 9 in section A.12 or B.12 of the appendices A and B respectively), because it indicates how heavy a smoker is or was (smoking load).

The text 63 m was wrongly tagged and normalized meaning a DURATION of 63 minutes. However, the context indicates the text is part of the height of a patient (1,63 m), meaning 1 meter and 63 centimeters. The HeidelTime rule made this extraction because there was a white-space between the number 1 and the number 63.

- d) "feminino, 75 anos acompanhamento da diabetes e da creatinina parcial de urina : nitrito positivo, leucocitos 57, cor castanha, traços de proteínas creatinina 1 . 9 ckd/epi 25 ml/min hb glicada 8, 5 % glicose durante o dia : de manhã controlada e no início e fim da noite sempre mais alta hmf : dm 2 e has há 30 anos dislipidemia doença renal estagio 4 obesidade e sedentarismo cx previas : platina no braço e perna em uso de : losartana 30 mg 2x dia sinvastatina 20 mg 1 x dia omeprazol 20 mg 1 x ao dia hidroclorotiazida 25 mg 1 x dia fluoxetina 20 mg 1 x dia insulina : nph 50u de manhã e 30u de noite, regular apenas de meio dia quando esta alta, usa 5 u uso de metformina apenas quando glicose maior que 300 pa : 117/64 mmhg fc : 67 bpm mmii edema 1+/4+ peso : 108 kg diurese mais de 3 x ao dia evacuações : 1 x na semana conduta : retirar metformina nph 50 de manhã, 10 no almoço e 30u de noite atb com macrodantina 100 mg 2x ao dia"

- 1) Snippet: "traços de proteínas creatinina 1 . 9 ckd/epi [Chronic Kidney Disease Epidemiology Collaboration] 25 ml/min hb glicada [hemoglobina glicada] 8, 5 % glicose durante o dia"
- 2) Free translation: "traits of creatinine proteins 1.9 ckd / epi [Chronic Kidney Disease Epidemiology Collaboration] 25 ml / min hb glycogen [glycated hemoglobin] 8.5 % glucose during the day"
- 3) HeidelTime: "type": "DATE", "value": "XXXX-09-1", "text": "1 . 9"
- 4) Discussion: The text "1 . 9" was tagged, because a rule was implemented to accept the dot as a separator between a day and a month, like in 1.9, 1/9 or 1-9. It was added to the rule, because samples of this kind of separator were found in clinical texts written in Portuguese.

- e) "lucas, 69 anos, acompanhamento dm 2 . diagnostico de dm 2 há 16 anos - iniciou o uso da insulina há 10 anos - esta bem adaptada a droga . nega poliúria, polidipsia, polifagia . nega outras queixas . medicamentos : sinvastatina 20 mg - um cp a noite, espirolactona 25 mg : 1 cp pela manhã, furosemida - 40 mg, aas - 100 mg - 1 cp

apos o almoço, insulina nph : 30 u pela manha e 30 u a noite, levotiroxina 100 mcg : 1 cp pela manha, glibenclamida 5 mg : 1 cp apos cafe, almoço e jantar, metformina 850 : 1 cp apos cafe, almoço e jantar, losartana 50 mg (1 cp pela manha e 1 cp a noite) . exames lab (18/08) : potassio : 4, 8; ct : 184; hdl : 52; ldl : 90; cr : 0, 9; glicose : 159, hemoglobina glicosilada : 8, 17%; tsh : 5, 61; microalbuminuria 24 horas : 15, 6; transaminase piruvica : 37; transaminase oxalacetica : 23 . obesidade grau iii; dm 2 ."

- 1) Snippet: "microalbuminuria 24 horas : 15, 6; transaminase piruvica : 37;"
 - 2) Free translation: "microalbuminuria 24 hours: 15.6; pyruvic transaminase: 37;"
 - 3) HeidelTime: "type": "TIME", "value": "24:00", "mod": "APPROX", "text": "24 horas"
 - 4) Discussion: Normally the text 24 horas is a TE which indicates a DURATION of 24 hours. However, in this case, the word microalbuminuria in the given context is part of the name of a laboratory exam. The whole name of the exam is "microalbuminuria 24 horas".
- f) "56 anos paciente comparece encaminhada devido a identificação de imagem polipoide em colo uterino . nega sintomatologia . nega sinusiorragia . nega sangramentos apos a menopausa . menopausa: aos 45 anos g 2 c 2 a 0 nega uso de th . amamentou os filhos ao seio cerca de 1-2 anos . am: nega patologias . af: nega historia familiar de neoplasia . nao traz exames complementares . ao exame: pa 120x80 peso 74, 9 ca 94 vulva hipotrofica, sem lesões . colo epitelizado, com polipo endocervical visualizado pelo orificio externo do colo, com pediculo parecendo vir do interior cavidade uterina . realizado retirada por torção do pediculo do polipo endocervical . revisão de hemostasia . material enviado para ap ao toque utero de volume normal, movel, regular . anexos nao palpaveis . cd: retorno em 45 dias . oriento paciente . doutor vital brazil r3"
- 1) Snippet: "menopausa: aos 45 anos g 2 c 2 a 0 nega uso de th [terapia hormonal]."
 - 2) Free translation: "menopause: at 45 years g 2 c 2 a 0 denies th use [hormonal therapy]."
 - 3) HeidelTime: "type": "DURATION", "value": "P2Y", "text": "2 a"
 - 4) Discussion: HeidelTime tagged the text 2 a, as if it meant a DURATION of 2 years. However, it is not correct, because text "g 2 c 2 a 0" mean: g 2 (gravidez 2 - two pregnancies) c 2 (cesarianas 2 - two cesarean sections) a 0 (abortos 0 - zero abortions). The letter "a" is the abbreation for the word "ano" (year).
- g) "primeira consulta irc - nefroesclerose hipertensiva? com gesf secundária? tgf estimada em 20 hematuria - com história de tabagismo lucas, 53 anos - paciente

encaminhada por piora de função renal - queixa de astenia - portadora de has com diagnóstico há cerca de 4 anos, com controle regular - hipotireoidismo com dx há um ano . - nega cirurgias prévias - itus prévias uso de: losartan 50mg 12/12h, furosemida 40mg, levotiroxina 100mcg, omeprazol 20mg . - ex tabagista 40 anos maço, parou há 5 anos - sedentária - nega história de doença renal na família trouxe exames laboratoriais: - hb 12 - creatinina 2, 7 (tfg estimada 20ml/min) - hematuria - us - rins 9, 2 e 9, 4, com cistos simples (3 no total) , bexiga normal - 2012 - proteinuria de 500mg ao exame: corada, hidratada, afebril pa 140/90 mv positivo, simetrico abdomen sp"

- 1) Snippet: "com gesf [glomerulosclerose segmentar focal] secundária? tgf [O correto seria tfg que significa Taxa de Filtração Glomerular] estimada em 20 hematuria - com história de tabagismo"
 - 2) Free translation: "with gesf [secondary focal segmental glomerulosclerosis]? tgf [The correct would be tfg which stands for Estimated Glomerular Filtration Rate] estimated in 20 hematuria - with a history of smoking"
 - 3) HeidelTime: "type": "DATE", "value": "2020", "text": "em 20"
 - 4) Discussion: In this case, the rule created to extract a year typed with two digits, like "em 80", meaning the year 1980, was incorrectly extracted, as "em 20" was not a year, but it was a percentage.
- h) "43 anos has desde os 17 anos, dm tipo 2, há 5 anos g 4 c 4 a0 laqueadura cateterismo cardíaco há 2 anos, em ponta grossa . sem indicação angioplastia arritmia dlp mista ex tabagista, , parou há 6 meses (fumava desde 9 anos) nega alteração tiróide, iam, ave em uso de lst 50 2x, aas, atenolol 50 2x, furosemida 40, metformina 850 hoje peso 82 pa 170/100 mucosas hipocoradas cpp livres abdome sp membros edema +/4 . suspenso metformina aumento furosemida e associao calcio como quelante . ecografia 28/05/2014 > sinais de nefropatia cronica bilateral 15/04/2014>: cr 3, 2 31/03/2015: cr 2, 47, cl cr 42 12/06/2015: glico 79, na 140, ca 9, 5, ac urico 6, 5, cr 2, 92, k 5, 2, ureia 98, p 5, 5, ht 41, hb 14, ph 7, 23, bic 21 p urina l 13mil, prot 100 a # drc > mdrd 28 acidose : metformina? anemia discreta caxp 52, 5 ."
- 1) Snippet: "ureia 98, p [potássio] 5, 5, ht [hematócrito] 41, hb [hemoglobina] 14, ph [potencial hidrogeniônico] 7, 23, bic [bicarbonato] 21 p [parcial] urina l 13mil, prot [proteinúria] 100 a # drc > mdrd [doença renaç crônica > Modification of Diet in Renal Disease] 28 acidose : metformina? anemia discreta caxp 52, 5 ."
 - 2) Free translation: "urea 98, p [potassium] 5, 5, ht [hematocrit] 41, hb [hemoglobin] 14, ph [hydrogenation potential] 7, 23, bic [bicarbonate] 21 p [partial] urine

13mil, prot [proteinuria] 100 a # drc > mdrd [chronic renaç disease > Modification of Diet in Renal Disease] 28 acidosis: metformin? Discrete anemia capx 52, 5."

3) HeidelTime: "type": "DURATION", "value": "P100Y", "text": "100 a"

4) Discussion: The text 100 a was tagged as if it meant a DURATION of "100 anos" (100 years), what is not correct according to the context where the tagged text is inserted.

i) "lucas, 62 anos comorbidades: # mcp dilatada / chagas # has # dislipidemia # marca-passo há 5 anos # dac familiar chv: ex-tabagista de 38 maços-ano hma: # paciente com icc classe funcional ii retorna com resultado do cat solicitado na última consulta; relata melhora do edema desde a última consulta/ cat: do dia 29/07/2014: protocolo pré - transplante cardíaco realizado com sucesso; rvp wood positiva: candidato atransplante; medicamentos: # enalapril 10 mg; 1/2 cp de 12/12 horas # apresolina 25 mg 1 cp de 8/8 horas # monocordl 20 mg 1 cp as 8 e outro as 14 horas # espironolactona 25 mg 1 cp ao dia # furosemida 40 mg 1 cp as 8 e outro as 14 horas # carvedilol 25 mg 1 cp de 12/12 horas # sinvastatina 20 mg 1 cp a noite conduta: solicitado internamento com monitorização para realização de transplante ;"

1) Snippet: "enalapril 10 mg; 1/2 cp de 12/12 horas"

2) Free translation: "enalapril 10 milligrams; 1/2 capsule 12/12 hours"

3) HeidelTime: "type": "DATE", "value": "XXXX-02-1", "text": "1/2"

4) Discussion: In this case, the text 1/2 was tagged by the tagger as meaning 1st of February. However, it is not correct, because it means half tablet.

j) "paciente em primeira consulta refere bav para perto há anos longe há anos refere dm tipo 1 em tto descontrolada há 8 anos nega has hf de cegueira e glaucoma nega cx e trauma ocular ao av: 20 / 70// 20 / 200 esq: -1, 00 esf ao tol: -1, 25 esf vv 20 / 40 // -0, 75 esf vv 20 / 70 fo: papila e macula nromais ao sem outras alterações evidentes ao pio: 14/13 bio: ca calma e bem formada com pupilas fotoreagnmets bilateral e simetricamente ao mr ao e bio de fundo sem alterações evidentes ao cdta: solicito angio e quando conseguir data para exame bater na porta para marcar retorno r1 doutor vital brazil r2 doutor vital brazil"

1) Snippet: "papila e macula nromais ao sem outras alterações evidentes ao pio: 14/13 bio: ca calma [nenhum dos profissional da saúde envolvidos nesta pesquisa soube o significado dessas abreviações **pio: 14/13 bio: ca calma**]"

2) Free translation: "normal papilla and macula without other obvious changes to the pio: 14/13 bio: ca calm [none of the health professionals involved in this research knew the meaning of these abbreviations **pio: 14/13 bio: ca calm**]"

- 3) HeidelTime: 'type': 'SET', 'value': 'P1W', 'text': 'ao sem'
- 4) Discussion: The text ao sem was tagged because a rule to tag the text **por sem** (por semana - per week) was created. The word "por" can be replaced by the word "ao" in the regular expression. This is the cause of this False Positive in the extraction step.

III. The following examples are cases that HeidelTime failed to extract TEs that existed in the test set (False Negatives).

- a) "lucas 28 anos gnc, perda renal pós-gravidez, bx realizada > não sabe laudo has drc, em trs - hd junho 2015 nega dm, dlp, tiróide, cardiopatias, litíase, outros . . . g 2 a 1 c 1 . nega outras cirurgias, diurese de bom volume . em uso de renalvit, epo, calciolit . internações > últimas por infecção permcath . trocado **em nov** huc e retirado **em dez** por febre . hoje com dl em jid, sem queixas há mais de 1 m, ês . peso seco 66, alt 160 . ao exame : pa 150/90, mucosas hipocoradas, demais exame fisico normal . liberada para transplante, esclareço riscos e beneficios, abo o, sorologias cmv e ebv reagente, antihbs reagente, toxo nr, ecocardiograma normal, doppler pelvico normal, ecografia abdome total > rins nefropatia, demais sp, ecg sinusal, fc 113, normal, te ausencia alterações isquemicas, imuno : rever mm e pra na reunião ."

- 1) Snippet: "trocado **em nov** huc [hospital] e retirado **em dez** por febre ."
- 2) Free translation: "changed **in nov** [November] huc [hospital] and withdrawn **in dec** [December] because of fever."
- 3) Test set: "text": "em nov", "type": "DATE", "mod": "NA", "value": "2015-11"
- 4) Test set: "text": "em dez", "mod": "NA", "type": "DATE", "value": "2015-12"
- 5) Discussion: The TEs **em nov** (in November) and **em dez** (in December) did not appear in the training set, it is why they did not became a rule in the HeidelTime we customized to the clinical texts in Brazilian Portuguese. The preposition "em" was added by the annotators, because it was a way to explicit that it is a DATE type, instead of a DURATION type, which is represented only by the name or the abbreviation of the name of the month.

- b) "lucas, 52 anos ic isquemica (rvm em 2008, atc em 2009 e iam + pcr + atc da em 2011) deficit neurológico pós pcr em 2011 (perda memória recente) em uso de: enalapril 10mg 12/12hs, carvedilol 25mg 12/12hs, espironolactona 25mg / d, aas 200mg / d, monocordil 20mg 2x, sinvastatina 40mg / d, levotiroxina 50**mcg** / **d** +25**mcg** / **dia**, quetruse lamitor e exodus trouxe lab hb 13, 1, leuc 7730, na 141, k 4, 3, craet 0, 8, , glic 94, ureia 72, cpk 126, tgp 27, tgo 23, ct 101, tsh 3, 44, hb glicada 5, 87, tg 120, hdl 28, ldl 50 ecocardio 028/05/16 ae 44, vd 23, ve 52/38, feve

52% ve hipertrofiacdo . acinesia medio basal infero dorsal, alteracao relaxamento s # exceto pela perdida memoria . refer estar assintomatico r sem estase jugular ac: bcrnf sem sopros b2 desdobrada fixa mv+ e simétrico, sem ra . abd: sp mmii: sem edema boa perfusão periférica a # ic isquemica cf ii - e remodelamento reverso p # suspendo monocordil e espirolactona . reduzo aas para 100mg solícito lab + ecott de controle retorno em 3 meses"

- 1) Snippet: "levotiroxina 50mcg / d +25mcg / dia"
 - 2) Free translation: "levothyroxine 50 **micrograms** / **day** +25 **micrograms** / **day**"
 - 3) Test set: "text": "mcg / d", "mod": "NA", "type": "SET", "value": "P1D"
 - 4) Test set: "text": "mcg / dia", "type": "SET", "mod": "NA", "value": "P1D"
 - 5) Discussion: The TEs **mcg / d** and **mcg / dia** were not present in the training set, this is the motive why they were not tagged by the tagger. The cause of these false negatives is that the abbreviation "mcg" was not present in the training set. However, the abbreviation "mg" was made part of a rule in the customized version of HeidelTime that we made.
- c) "lucas, 57 anos # dm, # has # dac - iam em 2009 -> atc . # rvm em 2009 . # atc em 2013 de cd + implante de mp (bav de 2 grau mobitz 2) # dislipidemia . # tabagista há mais de 40 anos (10 cigarros /dia) . # em uso de insulina nph 26ui / 10ui, metformina 850mg **café / almoço / janta**, aradois 50mg /dia, anlodipina 5mg 12/12 hs, enalapril 20mg 12/12 hs, sinvastatina 80mg, selozok 50mg /dia, aas 100mg, s # nega precordialgia, nega dispneia, ortopneia, dpn . refere claudicação intermitente para moderados esforços . o # beg, lote, corado, hidratado, eupneico, afebril, pa 150/80, fc 74, cp: sp, acv: bcrnf 2t ss, ap: mv+ reduzido difusamente, sem ra, abd: sp, mmii: pulsos reduzidos bilat, sem edema ou empastamento, lab 13/01/14: glicose 304; glicose pós-prandial 309; ct 119; hdl 21; tg 214; tgo 19; hba1c 6, 70; cpk 72; cr 1, 00; k 4, 7; ur 30; microalbuminuria 24 hs (2114 mg) ; pu (glicose ++; leuc 2) . a # dac, has, dm . p # aumento selozok para 100mg /dia . demais mantido . r3 cardio lucas, dr vital brasil ."

- 1) Snippet: "metformina 850mg **café / almoço / janta**"
- 2) Free translation: "metformin 850 milligrams at **breakfast / lunch / dinner**"
- 3) Test set: "text": "café / almoço / janta", "type": "SET", "mod": "NA", "value": "P1D", "freq": "3X"
- 4) Discussion: A pattern of the type SET which was not added to the custom HeidelTime rules, because it also did not existed in the training set.

d) "lucas, 29 anos s # segunda consulta . paciente veio acompanhado da mãe . refere melhora do tremor de mmss com a retirada do haldol, no entanto refere piora da acatisia e discinesia com a introdução da risperidona 02 mg (**0-0-1**) . refere sono não reparador, insônia intermediária . nega sintomas de hiperativação, nega alterações de sensopercepção, nega sintomas depressivos . mantém sintomas negativos - lentificação psicomotora e de pensamento, embotamento afetivo, apatia . trouxe exames lab: lítemia 0, 4 (nao tomou remédio na noite anterior) . em uso: risperidona 02 mg (**0-0-1**) citalopram 20 mg (**1-0-0**) carbolitium 300 mg (**1-0-2**) clonazepam 0, 5 mg (**0-0-1**) a # tab? personalidade esquizoide? ? p # conforme orientações doutor vital brazil mantenho prescrição médica, aumento clonazepam 0, 5 para (**2-0-1**) para melhora de acatisia e discinesia . retorno em 30 dias . o # aseado, cooperativo, orientado alo e auto, normovigil, normotenaz, humor hipotímico afeto embotado, crítica e insight parciais ."

- 1) Snippet: "discinesia com a introdução da risperidona 02 mg (**0-0-1**)"
- 2) Free translation: "dyskinesia with the introduction of risperidone 02 milligrams (**0-0-1**)"
- 3) Test set: "text": "0-0-1", "mod": "NA", "type": "SET", "value": "P1D"
- 4) Snippet: "em uso: risperidona 02 mg (**0-0-1**) citalopram 20 mg (**1-0-0**) carbolitium 300 mg (**1-0-2**) clonazepam 0, 5 mg (**0-0-1**) a # tab? personalidade esquizoide? ? p"
- 5) Free translation: "in use: risperidone 02 mg (**0-0-1**) citalopram 20 mg (**1-0-0**) carbolitium 300 mg (**1-0-2**) clonazepam 0.5 mg (**0-0-1**) a # tab? schizoid personality? ? P"
- 6) Test set: "'text': '0-0-1', 'type': 'SET', 'mod': 'NA', 'value': 'P1D'"
- 7) Test set: "'text': '1-0-0', 'mod': 'NA', 'type': 'SET', 'value': 'P1D'"
- 8) Test set: "'text': '1-0-2', 'mod': 'NA', 'type': 'SET', 'value': 'P1D', 'freq': '2X'"
- 9) Test set: "'text': '0-0-1', 'mod': 'NA', 'type': 'SET', 'value': 'P1D'"
- 10) Snippet: "aumento clonazepam 0, 5 para (**2-0-1**) para melhora de acatisia e discinesia ."
- 11) Free translation: "increase clonazepam 0.5 to (**2-0-1**) for improvement of akathisia and dyskinesia."
- 12) Test set: "'text': '2-0-1', 'mod': 'NA', 'type': 'SET', 'value': 'P1D', 'freq': '2X'"
- 13) Discussion: All TEs presented in this item did not exist in the training set. They follow a pattern that indicates how many times a medicine has to be taken per day. The pattern also indicates the quantity of medicine that has to be taken each time it is taken. For example, the pattern "1-0-2" indicates the

patient has to take one tablet in the morning and two tablets in the evening. The pattern 1-1-1 indicates the patient has to take one tablet in the morning or at breakfast, one tablet at lunchtime and one tablet in the evening. This pattern can also be understood as: take a tablet thrice a day or at each eight hours.

e) "id # lucas, 60 anos . qp # sem queixas . hma # nega dispnea, palpitações, dor toracica ou outros sintomas cardíacos . edema matinal . acompanha no ambulatório por cardiomiopatia dilatada idiopatica e fibrilação atrial . comorbidades: has, dm e gota . atualmente em uso furosemida 40mg 1 cp ao dia, digoxina 0, 25mg 1 cp/dia, carvedilol 25mg 2 cp/dia, monocordil 20mg 2 cp/dia, hidralazina 25mg 2 cp/dia, alopurinol 300mg 1 cp/dia, colchicina 0, 5mg se dor, marevan 5mg 2 cp seg, qua e sex, sab e **1 cp nos demais**, metformina 850mg 1 cp após almoço . hmp # já estive na fila para transplante - mretirado por melhora clinica . cate há aprox 2 anos - sem evidência de doença coronariana . internado por insuficiência renal? ?? (inicio **2014**) . herniorrafia de parede abdominal (sangramento retroperitoneal por discrasia sanguinea) há 1 ano . chv # nega tabagismo e etilismo, sedentário . hmf # mãe falecida por trombose aos 64 anos e pai falecida nefropatia com 78 anos; filhos saudáveis . exames: leucocitose discreta - 10800; - segmentados 65, 4; ureia: 92; cr: 2, 1; acido urico: 7, 3; sodio: 138; k: 5, 7; tfg (ckd-epi) : 33 ."

- 1) Snippet: "marevan 5mg 2 cp seg, qua e sex, sab e **1 cp nos demais**, metformina 850mg 1 cp após almoço ."
- 2) Free translation: "marevan 5 milligrams 2 capsule Mon, Wed and Fri, Sat and **1 capsule in the others**, Metformin 850 milligrams 1 capsule after lunch."
- 3) Test set: 'text': '1 cp nos demais', 'mod': 'NA', 'type': 'SET', 'value': 'P1W', 'freq': '3X'
- 4) Snippet: "internado por insuficiência renal? ?? (inicio **2014**) ."
- 5) Free translation: "hospitalized for kidney failure? ??? (beginning **2014**)."
- 6) Test set: 'text': '2014', 'mod': 'NA', 'type': 'DATE', 'value': '2014'
- 7) Discussion: The TE **1 cp nos demais** is another example of a pattern that was not found in the training set and, consequently, did not become a rule in the rules of HeidelTime we created or altered.

The TE **2014** was not tagged because of the left parenthesis ")" which was beside the TE. This was a limitation we found in HeidelTime, it used to appear in cases we were not able to track.

f) "lucas, 60 anos # ic etiologia isquemica # iam em 27/08/2010 # pot atc c stent # pot rvm 2011 mie - da / sf - diag # em uso de: - carvedilol 25mg 12/12hs -"

rosuva 10 - valsartan 320 - furo 40 - apresolina 50mg 12/12hs - espiro 25 - aas 100 - monoc **20mg noite** - mtf 850 3x - levotirox 50 - indapamida 1, 5 s # relata dispneia para moderados esforços . nega dor precordial, edema de mmii, ortopneia e dpn . - cintilo (06/10/14) : hipoperfusão persistente (fibrose) de grande extensão, no segmento medioapical da parede anterior e ápice de ve, em grau discreto na parede anterosseptal . fe repouso 49% - eco 17/09/15: ae 42, vd 26, se 10, pp 9, dd 55, ds 46, fe 34%, psap 49 mmhg . ve acinesia antero-septal e anterior medio-apical e apical . - lab : 23/04/15: k 5, 1, cr 1, 03, t4 livre 2, 26 (1, 74) o # pa 100/60, fc 72 sat 98% corado, hidratado, eupneico bcrnf ss 2t mv+ sim bilart sra abd sp mmii s edema a # ic isquemica p # mantenho medicações - solicito novos exames laboratoriais (avaliar função da tireoide - t4 acima do valor de referencia e função renal k 5, 1) - demais mantido - retorno em 4 meses"

- 1) Snippet: "espiro 25 - aas [Ácido acetilsalicílico] 100 - monoc [monocordil] **20mg noite** - mtf [metformina] 850 3x "
 - 2) Free translation: "spiro 25-aas [Acetylsalicylic acid] 100-monoc [monocordyl] **20mg night** -mtf [metformin] 850 3x"
 - 3) Test set: 'text': '20mg noite', 'type': 'SET', 'mod': 'NA', 'value': 'P1D'
 - 4) Discussion: The TE **20mg noite** was not extracted because the Heideltime rule we made did not considered the missing white space between the number 20 and the abbreviation. It is an example of a noisy TE.
- g) "lucas 60 anos comorbidades: iam há 5 anos - cateterização, icc, dislipidemia, has, dm muc: aas 100mg 1cp, enalapril 10mg 1cp 12/12hrs, selozok 100mg 1cp 12/12hrs, espirolactona 25mg 1cp, sinvastatina 20mg 2cp, fluoxetina 10mg 1cp glifage 500mg **1cp noite**, insulina nph 28u m 24u n, insulina regular 12u n # s - em acompanhamento ambulatorial por dm tipo 2 há 5 anos, dx de dm há 23 anos - paciente vai fazer o laser para correção da visão - nega alterações de habito urinario e intestinal - não faz acompanhamento nutricional - pratica caminhadas de 1 hora pela manha e a atarde todo dia # o - beg, lote, anicterica, acianotica, corada e hidratada - cv : brnf 2t ss - cpp : spp - abs : spp - mmii : pes bem cuidados, com trofismo preservado, sem evidencia de onicomiose exames laboratoriais - cr 1, 2 tfg ckd epi 48 . 9 ml/min/1 . 73 m2- glicose 155 - hbac1 7, 86% - microalbuminuria 24 hrs 10, 8mg / 24hrs # a - dm tipo 2 em acompanhamento - sem evidencia de irc # p - paciente orientada quando ao controle da glicemia capilar - orientação para pega fitas e fazer cadastro em curitiba"
- 1) Snippet: "fluoxetina 10mg 1cp glifage [medicamento antidiabético] 500mg **1cp noite**, insulina nph [Insulina Humulin Neutra] 28u [unidade] m [manhã] 24u [unidade] n [noite]"

- 2) Free translation: "fluoxetine 10mg 1cp glifage [antidiabetic medicine] 500mg **1cp night**, insulin nph [Insulin Humulin Neutral] 28u [unit] m [morning] 24u [unit] n [night] "
 - 3) Test set: 'text': '1cp noite', 'mod': 'NA', 'type': 'SET', 'value': 'PID'
 - 4) Discussion: The TE **1cp noite** was not extracted because the HeidelTime rule we made did not considered the missing white space between the number 20 and the abbreviation. It is an example of a noisy TE.
- h) "lucas cmd após gestação . implantou ressinc e cdi em **nov de 15** . não realizou consulta em nosso ambu do mp . hoje veio como procura direta - encaixe . sente-se melhor porem não esta 100% . antes sentia-se mal até em pentear o cabelo, tomar banho (estava praticamente em cadeira de rodas) , tinha dpn e ortopneia . agora consegue caminhar empurrando carrinho por meia quadra . # mp multissitio . # programado em biv 70 70 . # operando mais de 95% em biv . # ritmo proprio = bradi sinusal . # a - imp = 537, sense = 2, 6, cla = 1, 2 x 0, 5 . # v - imp = 522, sense = 17, clv = 0, 7 x 0, 5 . ve com 609, 19, 6 e 11 x 0, 5 . # bat ok . # programado em fc de 60 - 50, mantido demais, retorno em 30 dias + não faltar consulta com doutor vital brazil ."
- 1) Snippet: "lucas cmd após gestação . implantou ressinc e cdi em **nov de 15** . não realizou consulta em nosso ambu [ambulatório] do mp . [nenhum dos profissional da saúde envolvidos nesta pesquisa soube o significado da abreviação **mp**]"
 - 2) Free translation: "lucas cmd after gestation. deployed ressinc and cdi in **Nov 15**. did not perform consultation in our ambu [outpatient] of the mp. [none of the health professionals involved in this research knew the meaning of the abbreviation **mp**]"
 - 3) Test set: 'text': 'nov de 15', 'mod': 'NA', 'type': 'DATE', 'value': '2015-11'
 - 4) Discussion: This is an example of a TE (**nov de 15**) that did not exist in the training set and, consequently, the pattern did not become a rule in HeidelTime.
- i) "nefro geral . lucas, 78 anos . has, em uso de enl **20 x2**, atn **50 x2**, hctz . hpb, episodio de obstrução urinária há 6 anos . nega cirurgias abdominais, ex-tabagista (parou há 30 anos) . avc em 2013, sem sequelas . exames agosto/2014: ph 7, 29; bic 22; pu 1, 015; leuc 6m, hem 1m, gj 93, na 143, ct 165, cr 2, 02; k 5, 8, ur 101, ca 9, 2, fe 91, p 3, 6, hdl 23, tgl 161, acido urico 9, 4, ldl 110, uc neg . rx com ectasia da aorta . us: cistos renais simples bilaterais, espessura cortical normal, sem litíase, sem dilatação, próstata 44g, com resíduo vesical pós-miccional (139ml) . ao exame: 147 cm, 69 kg, imc 31 . corado, hidratado, pa 150 x 70 (msd e mse) , fc 80, crepitanes bibasais, abdomen globoso sem vcm, edema periférico ."

- 1) Snippet: "nefro geral . lucas, 78 anos . has [hipertensão arterial sistêmica], em uso de enl **20 x2**, atn **50 x2**, hctz . hpb [hiperplasia prostática benigna], episodio de obstrução urinária há 6 anos ."
 - 2) Free translation: "general nephrology. lucas, 78 years old. has [systemic arterial hypertension], in use of enl **20 x2**, atn **50 x2**, hctz. hpb [benign prostatic hyperplasia], episode of urinary obstruction 6 years ago"
 - 3) Test set: 'text': '20 x2', 'mod': 'NA', 'type': 'SET', 'value': 'P1D', 'freq': '2X'
 - 4) Test set: 'text': '50 x2', 'mod': 'NA', 'type': 'SET', 'value': 'P1D', 'freq': '2X'
 - 5) Discussion: The TEs **20 x2** and **50 x2** are also examples of TEs that did not exist in the training set and, did not become a rule in HeidelTime.
- j) "# **12º** po bx lesao mandibula à direita + bx perfuração nasal (ver ap) . s # fez uso de antibiotico (amoxi+ clav) e corticoide oral com melhora importante dos sintomas . refere redução quase completa do edema endurecido em mandibula à direita . # hmp: transplante renal à direita há 5 anos - em acompanhamento com nefrologia . o # oro: fo bom aspecto . c/p: edema endurecido em mandibula aprox 5cm . rino: perf nasal anterior grande crostas c/ crostas ao redor . cd # retorno em 2 semanas para ver ap! (consulta em 1 semana coincidiria com consulta da nefro) . orientações gerais . manutenção de medicação analgésica se dor . r2 lucas"
- 1) Snippet: "# **12º** po bx lesao mandibula à direita + bx perfuração nasal (ver ap [anatomopatológico]) ."
 - 2) Free translation: "# **12º** [12th day] po bx right jaw lesion + bx nasal perforation (see ap [pathological anatomy])."
 - 3) Test set: 'text': '12º', 'type': 'DATE', 'mod': 'NA', 'value': 'XXXX-XX-XX'
 - 4) Discussion: The TE **12º** is another example of a pattern which did not appear in the training set and, was not added to the HeidelTime rules we created.

IV. The examples below present cases of False Positives in the normalization step made by HeidelTime.

- a) "lucas, 68 anos . cat em 2008 . valvoplastia aórtica em 2008 . valva biológica . 40 dias de arritmia . cardioversao, ficou bom . queixa de cansaço antes da cx - pneumonia (sic) . fumante por 40 anos, 25 cigarros por dia . refere estar sem sintomas cardiovasc . ultima consulta há 1 ano . pa baixa (sic) . em uso de: - enalapril 10 - 12/12; - carvedilol 6, 25 - 12/12; - aas 100; - atorvastatina 20 - 2 a noite . exame fisico: fc 80, pa 150/100, fr 16; bernf com sopro sistólico em foco aortico, mv+ bilateral, sem ra . rh+, sem visceromegalia ou dor a palpação . hernia umbulical, anel herniário de 5cm, redutível ."

- 1) Snippet: "carvedilol 6, 25 - 12/12; - aas 100;"
 - 2) Free translation: "carvedilol 6, 25 - 12/12; - aas 100;"
 - 3) Test set: type:"SET" value:"P1D" text:"12/12" freq:"2X"
HeidelTime: type:"DATE" value:"2012-12-25" text:"25 - 12/12"
 - 4) Discussion: The text **12/12** can be easily confused with a date type. The extraction can be considered correct, from a relaxed matching point of view, because the same snippet is present in the test set. However, the normalization is incorrect, because the text **12/12** means: take the medicine each 12 hours, or twice a day.
- b) "g2p1, ig de 36 sem 6 dia (s) , ts o + . interna para realizar cesárea eletiva por indicação fetal (macrocrania + displasia óssea, com hipoplasia pulmonar importante) . parto cesárea em 26/06/2007 19:24 hs, rn apgar 1/2, masc, 3085 g . morte neonatal prematura por múltiplas malformações incompatíveis com a vida . pré-natal neste hospital com equipe freitas . exames hcpc : vdrl: nr anti-hcv: nr anti-hiv: nr hbsag: negativo toxo igg e igm: nr recebe alta em bom estado geral, sinais vitais estáveis, mamas lactantes, lóquios fisiológicos, útero contraído, ferida operatória limpa e seca ."
- 1) Snippet: "parto cesárea em 26/06/2007 19:24 hs, rn [recém nascido] apgar (é uma escala) 1/2, masc, 3085 g ."
 - 2) Free translation: "cesarean delivery" in 26/06/2007 19:24 hours, rn [new-born] apgar (it is a scale) 1/2, male, 3085 g."
 - 3) Test set: type:"TIME" value:"2007-06-26T19:24" text:"26/06/2007 19:24 hs"
HeidelTime: type:"DATE" value:"2007-06-26" text:"em 26/06/2007"
HeidelTime: type:"TIME" value:"19:24" text:"19:24 hs"
 - 4) Discussion: In this case, Heideltime split a single TE in two parts, what does not make sense, since it is just one TE. This pattern did not appear in the training set, consequently, a rule in the tagger was not create with this aim.
- c) "paciente portados de has, mutação do gene da protrombina em anticoagulação crônica com warfarin com iam e avc em 2001 com sequela (déficit motor à esquerda) com cat de 2001 evidenciando hipertrofia de ve e coronárias normais . ergometria de **06/04** com baixa capacidade funcional . em 30/12/04 iniciou com dor retroesternal, em aperto, intensa sem irradiação . veio à emergência do xxxx . ecg inalterado em relação aos prévios . marcadores de injúria miocárdica sem alteração e com alívio da dor com isordil sl . frente ao diagnóstico de angina instável, o paciente foi internado na uti coronariana e iniciado enoxaparina, propranolol e nitrato fixo . evoluiu com alguns episódios de dor e após 36h sem dor, sem alteração eletrocardiográfica

ou enzimática, recebeu alta para a internação na equipe de cardiologia em beg, hemodinamicamente estável, pa controlada, com warfarin reiniciado e com plano de suspender enoxaparina após 7 dias de warfarina (inr entre 2, 0 e 3, 0) . no andar evoluiu cardiologicamente bem . a enoxaparina foi suspensa quando o inr estava em 2, 38 . paciente recebe alta em beg, anticoagulado, sem sangramentos, hemodinamicamente estável, eupnéico, afebril e com plano de retorno ambulatorial (cardiologia) com o dr vital brasil . sai com prescrição domiciliar de : - propranolol 40 mg tid - aas 100 mg / dia - isossorbida 40 mg tid - isossorbida 5 mg sl s/n - varfarina 5 mg /dia ."

- 1) Snippet: "ergometria de **06/04** com baixa capacidade funcional ."
 - 2) Free translation: "ergometry of **06/04** with low functional capacity."
 - 3) Test set: type:"DATE" value:"2004-06" text:"06/04"
HeidelTime: type:"DATE" value:"XXXX-04-06" text:"06/04"
 - 4) Discussion: This is a common mistake made by the tagger in the normalization step. the TE **06/04** refers to a month/year, however, the tagger "understood" it as day/month. It is a hard confusion to avoid using a regular expression based tagger.
- d) "75 anos, natural de ibitinga-sp, residente em curitiba, trabalhou na roça, como tecelã e agora trabalha em casa, casada . qp: veio para acompanhamento . hma: paciente refere dm há 8 anos, toma metformina 850 mg 2x por dia . descobriu ao acaso durante um exame de sangue . paciente tem hipertireoidismo há mais de 5 anos descoberto em exame . toma tapazol 10mg 2 x/dia . paciente refere cansaço nos mmii final da tarde e nega edema . refere dor no braço e, há muitos anos que começou com piora há 2 anos e melhora com fisioterapia e paracetamol . mãe faleceu de ca no fígado e pai faleceu de pneumonia . tios tiveram ca de estômago ou fígado . has há 20 anos aproximadamente controlada com medicamentos . fez 3 cx de glaucoma e catarata, a última há 5 anos e faz acompanhamento com oftalmologista . fez apendicectomia com **14 anos** . pcte refere arritmia e distúrbio de coagulação acompanhado com cardiologista . g 8 p 8 c 0 a 0 . menopausa com **45 anos**, nega metrorragia . nega reposição hormonal . digoxina 0, 25mg . hidroclorotiazida 25mg . espironolactona 25mg . omeprazol 20mg . carvedilol 25mg . sinvastatina 20mg 2cp a noite . januvia 100mg 1cp no café . valsartana 320mg . marevan 5mg . metformina 850mg . tapazol 10mg . exames: creatinina 1, 2 . gj 113 . cpk 90 . calcio total: 8, 9 . tg 302 . hdl 24 . colesterol total 167 . ldl 83 . tsh 4, 07 . t4 0, 68 . vit d 36, 1 . hba1c 7, 30% . glicose pós-prandial 119 ."
- 1) Snippet: "fez apendicectomia com **14 anos** . pcte refere arritmia e distúrbio de coagulação acompanhado com cardiologista . g 8 p 8 c 0 a 0 . menopausa com

45 anos, nega metrorragia ."

- 2) Free translation: "had an appendectomy at the **age of 14**. arrhythmia and coagulation disorder accompanied by a cardiologist. pregnancy 8 (g 8), premature 8 (p 8), cesarean section 0 (c 0), abortions 0 (a 0). menopause at **age 45**, denies metrorrhagia."
 - 3) Test set: tipo:"DATE" normalização:"XXXX-XX-XX" text:"14 anos"
HeidelTime: tipo:"DURATION" normalização:"P14Y" text:"14 anos"
 - 4) Test set: tipo:"DATE" normalização:"XXXX-XX-XX" text:"45 anos"
HeidelTime: tipo:"DURATION" normalização:"P45Y" text:"45 anos"
 - 5) Discussion: In most of cases, there is an ambiguous interpretation about some TEs which can be classified as DATE or DURATION. However, in this case TEs cannot be normalized as DURATION according to the context of the clinical report. The confusion between types DATE and DURATION is also reported by [Hagège, Baptista and Mamede \(2008\)](#) and [Viani et al. \(2018\)](#).
- e) "lucas, 62 anos # eao # has # uso: atenol 50 - 1xd . enalapril 5 - 12/12h . # sem queixas cardiovasculares . ecocardio (13/02/15) : ae = 45; siv = ppve = 14mm; ddve = 48; dsve = 29; fe = 69% . vmi = folhetos espessados, abertura preservada, refluxo discreto, fluxo diastólico com padrão e<a vao = cúspides calcificadas, com dupla lesão; grad max de 54 e medio de 28mmhg . vp = ok; vtri = folhetos espessados, abertura preservada, insuficiente, psap = 39mmhg . ve = hipertrofiado, cavidade, função sistólica e contração preservadas, alt de relaxamento . ad = aumentado . eco (14/04/16) ra: 38 ae: 45 vd: 20 s: 14 pp: 14 ve: 50 / 30 fe: 70% ao: cuspides calcificadas, com dupla lesao . grad max = 43 medio 25 . insuf tricu discreta (psap 39) refluxo mitral discreto . ao asc: 41mm s . nega dispneia, dor toracica, sincope . nega edema de mmii . refere controle pressorico bom em casa, porem nas consultas sempre apresenta elevação de pa . sd jaleco branco ?) beg, eupneico, acianotico, anicterico pa: 170/100 mmhg, p = 99 spm mv normodist sem ruidos . rcr, 2t, taquicardico, ss ++/6+, crsec / decres, fao, irradiação para todo o precórdio . a # dupla lesao ao estenose mod + insuf discreta p # aumento enalapril 20mg 2x d e inicio hctz 25mg / dia . retorno em 1 ano com novo eco r4 cardio doutor vital brazil"
- 1) Snippet: "enalapril 5 - 12/12h . # sem queixas cardiovasculares ."
 - 2) Free translation: "enalapril 5 - 12 / 12 hours. # no cardiovascular complaints."
 - 3) Test set: type:"SET" value:"P1D" text:"12/12h" freq:"2X"
HeidelTime: type:"DATE" value:"XXXX-12-5" text:"5 - 12"
 - 4) Discussion: This is another example of confusion between a SET type which was normalized as DATE type.

f) "paciente interna com quadro de hemorragia digestiva alta, endoscopia de urgencia mostrou 2 ulceras, 1 com sangramento, colocado clip metalico . não apresentou mais sangramento, mas descompensou da angina, com muitos episódios de dor . hematócrito estava 9, 3, optou-se por transfundir 2 ui chad para melhora do controle dos sintomas . realizou endoscopia de controle **72 horas apos**, com resultado favorável das ulceras, em fase de cicatrização, sem evidencia de sangramento . alta em condições estáveis, com o mesmo perfil de angina dos ultimos meses . orientado para manter medicação prescrita e acompanhamento com equipe da dor e psiquiatria ."

- 1) Snippet: "realizou endoscopia de controle **72 horas apos**, com resultado favorável das ulceras,"
- 2) Free translation: "performed endoscopy of control **72 hours after**, with favorable result of ulcers,"
- 3) Test set: type:"DATE" value:"XXXX-XX-XX" text:"72 horas apos"
HeidelTime: type:"DURATION" value:"P3D" text:"72 horas"
- 4) Discussion: HeidelTime normalized the TE **72 horas apos** as DURATION, however, according to the context, the TE is a DATE type, as marked in the test set. It happened because this type of TE did not exist in the training set, which was used to create the clinical report domain rules. The confusion between types DATE and DURATION is also reported by [Hagège, Baptista and Mamede \(2008\)](#) and [Viani et al. \(2018\)](#).

g) "lucas, 70 anos # em acompanhamento por dscc e face s # paciente relata prurido e descamação em couro cabeludo . há 1 mês não usa o shampoo de cetoconazol está em uso de shampoo que compra na farmácia (clear anticaspas) . nega outras queixas . o # 1- pequenas placas eritemato descamativas em braços e mãos 2- pouca descamação em couro cabeludo . sem outras lesões suspeitas a # 1- ceratoses actínicas 2- dscc p # crio nas ceratoses actinicas de mãos e braços . therapsor solução 8 gotas no couro cabeludo **1 x por semana** cetoconazol shampoo 2% 3x semana oriento fps 3x dia . retorno em 4 meses . estagiária lucas doutor vital brazil"

- 1) Snippet: "therapsor solução 8 gotas no couro cabeludo **1 x por semana** cetoconazol shampoo 2% 3x semana oriento fps [acreditamos que seja "fisioterapia", mas não temos certeza em virtude do contexto.] 3x dia ."
- 2) Free translation: "therapsor solution 8 drops on the scalp **1 x per week** ketoconazole shampoo 2% 3x week guide fps [we believe it to be "physiotherapy", but we are not sure by virtue of the context.] 3x day. "
- 3) Test set: type:"SET" value:"P1W" text:"1 x por semana"
HeidelTime: type:"SET" value:"P1D" text:"1 x"

- 4) Discussion: "1 x" means in most of the cases "once a day", however in this case, it is not correct, because is "once a week".
- h) "lucas 60 anos comorbidades: iam há 5 anos - cateterização, icc, dislipidemia, has, dm muc: aas 100mg 1cp, enalapril 10mg 1cp 12/12hrs, selozok 100mg 1cp 12/12hrs, es-pironolactona 25mg 1cp, sinvastatina 20mg 2cp, fluoxetina 10mg 1cp glifage 500mg 1cp noite, insulina nph 28u m 24u n, insulina regular 12u n # s - em acompanhamento ambulatorial por dm tipo 2 há 5 anos, dx de dm há 23 anos - paciente vai fazer o laser para correção da visão - nega alterações de habito urinario e intestinal - não faz acompanhamento nutricional - pratica caminhadas de **1 hora pela manha e a tarde todo dia** # o - beg, lote, anicterica, acianotica, corada e hidratada - cv : brnf 2t ss - cpp : spp - abs : spp - mmii : pes bem cuidados, com trofismo preservado, sem evidencia de onicomiose exames laboratoriais - cr 1, 2 tfg ckd epi 48 . 9 ml/min/1 . 73 m2- glicose 155 - hbac1 7, 86% - microalbuminuria 24 hrs 10, 8mg / 24hrs # a - dm tipo 2 em acompanhamento - sem evidencia de irc # p - paciente orientada quando ao controle da glicemia capilar - orientação para pega fitas e fazer cadastro em curitiba"

- 1) Snippet: "pratica caminhadas de **1 hora pela manha e a tarde todo dia** # o - beg [bom estado geral], lote, anicterica, acianotica, corada e hidratada"
- 2) Free translation: "walks for **1 hour in the morning and in the afternoon every day** # o - beg [good general condition], lot, anicteric, aciantica, colored and moisturized "
- 3) Test set: type:"DURATION" value:"P1H" text:"1 hora"
HeidelTime: type:"TIME" value:"1:00" text:"1 hora"
- 4) Test set: tipo:"SET" normalização:"P1D" text:"pela manha e a tarde todo dia"
freq:"2X"
HeidelTime: type:"SET" value:"P1D" text:"pela manha"
HeidelTime: type:"SET" value:"P1D" text:"atarde"
HeidelTime: type:"DURATION" value:"P1D" text:"todo dia"
- 5) Discussion: The TE **1 hora** was tagged correctly by HeidelTime, however, the normalization was incorrect.

The TE **1 hora** was tagged correctly by HeidelTime, however, the normalization is incorrect.

The TE **pela manha e a tarde todo dia** from the test set was extracted and normalized wrongly by the tagger. The whole phrase should have been tagged and normalized as one TE. Nevertheless, HeidelTime split it in three parts **pela manha**, **atarde** and **todo dia**. The part **atarde** is an example of a noisy TE. This is another example of a TE which did not existed in the training set.

- i) "paciente com protese valvular mecanica aortica em 1995 por estenose aortica, re-operação em 1996 por leak paravalvular, em anticoagulação desde então, em acompanhamento com medico externo . veio à xxxx em 09/08/03 por dispneia + cansaço, apresentando bradicardia importante (fc em torno de 40 bpm) , ecg com bradicardia e bloqueio trifascicular . como usava digoxina, aguardou-se nível sérico para decisão sobre necessidade de marcapasso . em 14/08, entretanto, apresentou piora dos sintomas e foi então colocado mp definitivo . o paciente apresentou boa evolução após a colocação do mp, apenas com 2 episódios de epistaxe não volumosas, com inr = 2 . como deve persistir anticoagulado, foi realizada avaliação pela consultoria da otorrino, com lesão da mucosa septal por manipulação, sendo feito tto e será reavaliado ambulatorialmente . recebe alta em boas condições, com mp com bom funcionamento, retorno ambulatorial rpecoce com otorrino para reavaliação e retorno em **dezembro** para reavaliação do mp com cirurgia cardiaca . deve manter medicações em uso prévio e acompanhamento com seu médico assistente . "
- 1) Snippet: "retorno ambulatorial rpecoce com otorrino para reavaliação e retorno em **dezembro** para reavaliação do mp [marcapasso] com cirurgia cardiaca ."
 - 2) Free translation: "early outpatient return with otorrino for reassessment and return in **December** for review of MP [pacemaker] with cardiac surgery."
 - 3) Test set: type:"DATE" value:"2003-12" text:"dezembro"
HeidelTime: type:"DATE" value:"2014-12" text:"dezembro"
 - 4) Discussion: The year 2003 is calculated based on the date 09/08/03 found in the text. The year 2014 normalized by HeidelTime came from the previous clinical report, because we prepared the tagger to process clinical texts one by one in sequence, in a single file (with 20 clinical reports by file). The part that the wrong date might come from is "hipertrofia de trapézio . rm (325/07/2014) : angioma venoso na região parietal esquerda ." which belongs to the previous report number 10979193, in the same file. The chosen approach was not the best one, the correct way would be restart the tagger to work over each clinical text individually, not keeping any data from the previous clinical report in memory. Unfortunately, we have no time to make this adjustment. However, our method result was not favoured by this issue, by the contrary. If this issue did not happen, chances were higher that this TE, for example, would be normalized in the same ways as in the test set.
- j) "lucas, 5 anos paciente com epilepsia de difícil controle, rdnpm, em uso de topiramato e vpa, iniciou com crises convulsivas, além de queixa de febre, tosse e rinorréia . realizado rx tx: fc em lie . iniciado amoxicilina . em 20/06, iniciado beta-2 1/1h, o2 1l/min e ctc vo, por tempo expiratório aumentado e sibilância; à noite, espaçado beta-2 para de 2/2h . neuro avaliou caso; aumentou topiramato para 75mg 2x/dia

e manteve dose do valproato (impressão: escape? dose inadequada de anticonvulsivantes ?) . evoluiu afebril, com melhora do ponto de vista respiratório, sendo retirado o2 em 21/06 pela manhã . afebril há 24h . paciente mantém tempo expiratório levemente aumentado, saturando 95% em ar ambiente, nebulizando de 2/2h, caminhando, em bom estado geral . mãe insiste em ir embora e refere ter recursos para recorrer se piora do quadro . explico que seria mais prudente aguardar melhora completa . alta hospitalar com beta-2 de **3/3 horas**, etc vo e amoxicilina, além de anticonvulsivantes . retornar se necessário ."

- 1) Snippet: "alta hospitalar com beta-2 de **3/3 horas**, etc vo [cetoconazol via oral] e amoxicilina, além de anticonvulsivantes . retornar se necessário ."
- 2) Free translation: "hospital discharge with beta-2 of **3/3 hours**, etc orally [ketconazole] and amoxicillin, in addition to anticonvulsants . return if necessary ."
- 3) Test set: type:"SET" value:"P3H" text:"3/3 horas"
HeidelTime: type:"DATE" value:"XXXX-03-3" text:"3/3"
- 4) Discussion: The TE **3/3 horas** was wrongly normalized as DATE type by HeidelTime, instead of SET type, like in the test set.

Many situations involving extraction and normalization steps were presented in this chapter, whether they were true positive, false positive or false negative. Error cases like false positive and false negative can be solved using some strategies.

False positive in general can be solved improving the detailing level of the rules, altering them to be more specific to the error. However, instead of changing a rule to more specific, it may be better split the rule into other rules, keeping it able to match a more general case and use the new rules that derived from the general one to match more specific cases. As already mentioned, there are cases that more context is required, like the TE "03/12", meaning the rule has to match more words before and after the TE. But, this strategy makes rules weaker, in the sense that if a digit or anything different found in this portion of text the rule will not have a match. These cases usually are better dealt with machine learning approaches.

False negative error can be solved by creating rules to the cases not extracted by the tagger. In order to improve the extraction of noisy TEs, preprocessing alternatives can be applied, like using a string similarity approach, based on a predefined dictionary. The strategy used in our proposed method is also valid, which is to include noisy cases in the tagger rules, but it is a time consuming and expensive strategy, as usual in rule-based approaches.

The next chapter presents the conclusion of this work.

9 Conclusion

While time is a dimension of human existence, temporal expressions are found in all sorts of texts. Whether it is a news text or a clinical text. Temporal information can be useful for many purposes in computation, including Natural Language Processing. From a computational point of view, temporal expressions need to be extracted and transformed in a way that computers understand it. In other words, TEs must be extracted and undergo a normalization step before they can be used by machines. Once temporal information is "understood" by computers, it becomes useful in tasks such as: question answering, text classification, search and summarization, for example. While the most studied language concerning the time subject is English, news texts are the most studied domain. From a temporal information extraction perspective, there are three tasks related to time. They are temporal tagging, event tagging and temporal relation. Temporal tagging has two steps, extraction and normalization. Event tagging is the task of finding one or more words which are events related to time in a given text. The temporal relation task is related to how the linkage between a temporal expression and an event occur in a given text. The same or similar benefits of temporal information extraction in news domain texts can be found in clinical texts. Whether they come from Electronic Health Records (EHR) in English, France, Portuguese or any other language. A feature of clinical texts is the occurrence of noisy words, caused by the stressing work environment of health professionals. In this study, noisy words are understood as any kind of misspelling, syntactic error, semantic mistake or typing error. From a technical view, temporal tagging can be made through a rule-based approach, a machine learning approach or a hybrid approach.

This dissertation explores a temporal tagging method to extract and normalize incorrectly (non-standard) written temporal expressions (TEs) and correctly written ones from clinical texts written in Brazilian Portuguese. A rule-based temporal tagger called HeidelbergTime was used to accomplish this task. To deal with correctly and incorrectly written TEs, also called noisy TEs, three approaches were used. The first one was called "Correct TEs", the second one was called "Noisy TEs" and the last one was called "N-gram TEs". In the "Correct TEs" approach, the correctly written TEs found in the training set were transformed into rules of HeidelbergTime. They were appended with the rules for Portuguese that already existed in the tagger. In the second approach, the noisy TEs found in the training set were either transformed into rules of the tagger. It is important to mention that, the rules of the second approach were appended with the rules of the first approach. In the last approach called "N-gram TEs", the rules of the previous approaches were appended with rules created based in TEs found by a character based n-gram strategy. As already mentioned, rules creation followed an incremental approach. The statistical eval-

uation showed there was no difference between the last two approaches. However, there was a statistical difference when the "Correct TEs" approach was compared to the "Noisy TEs" approach. The latter achieved the best results. It reached a F1-Score of 0.9021 for the extraction step and a F1-Score of 0.9245 in the normalization step.

Considering the hypotheses established for this research, the first one reads "Noisy TEs from clinical text improve the assertiveness of a temporal tagger when the TEs are transformed into rules of this tagger" (H1). This hypothesis was accepted, as proven by the F1-Scores cited above, related to the "Noisy TEs" approach. This result consequently rejects the second hypothesis, which is "The creation of new rules based on parts (n-gram) of noisy TEs increase the quality of a temporal tagger" (H2). This hypothesis is related to the "N-gram TEs" approach, which did not show statistical difference with the "Noisy TEs" approach.

This research has also a secondary contribution, which is the creation of a real world corpus of clinical texts written in Brazilian Portuguese. We mean real world corpus in contrast with meticulously prepared corpora which are used in research competitions. This corpus contained 310 texts, where 190 texts were used as training set and 120 texts were used as test set. The corpus was labelled by four healthcare professionals, a senior nurse with a master degree and three medical students. The annotation process showed that the normalization step of real world clinical texts is a difficult task. This conclusion considers the low inter-annotator agreement value. The annotators felt challenged to interpret the attribute type (DATE, TIME, DURATION or SET) of clinical reports. The same understanding occurred in relation to the different ways they could assign the value attribute of a temporal expression, specially because of its ambiguity, where both options could make sense.

Our best result is on the order of other researches made for other languages, including clinical texts (GUPTA; JOSHI; BHATTACHARYYA, 2015; HAMON; GRABAR, 2014; MOHARASAR; HO, 2016), highlighting that they do not deal with noisy TEs in their extraction and normalization phases. The only work that made a very similar research to ours, coping with the same problem, domain and language achieve results inferior to ours (AZEVEDO et al., 2018).

A study on the same problem, domain and language using a machine learning approach or a hybrid approach, combining machine learning, string similarity algorithms and a rule-based approach merits further research. The results showed HeidelTime gets some TEs wrongly due to its limitation in considering the context around the TEs. This limitation might be overtaken by the use of machine learning approaches. The use of larger datasets may also help to improve the performance of any kind of approach, since building datasets for temporal tagging are time consuming and expensive.

Appendix

APPENDIX A – Guideline

Este é o guideline original, escrito em português, sua versão traduzida para o inglês está na sequência.

A.1 Motivação

O processo de marcação temporal é importante para sistemas computacionais cujo insumo é texto. Entre esses sistemas podemos elencar os destinados a question/answering, chatbots, criação de timeline e sumarização.

Para este trabalho assumimos o pressuposto fundamental de que: “Se um humano pode determinar um valor para uma expressão temporal, ela deve ser marcada” (FERRO et al., 2001). Em outras palavras, se uma pessoa entende que uma palavra ou conjunto de palavras denota tempo, essa expressão temporal deve ser marcada. Entretanto, é necessário que essa marcação também permita sua normalização. Nesse sentido, também é importante salientar que prevaleceram as expressões temporais que foram consideradas úteis pelos especialistas da área da saúde. Esses profissionais participaram da construção desse guideline, assim como também o utilizaram.

A.2 Expressão Temporal

A unidade básica para o processo de marcação temporal é a **Expressão Temporal (ET)**. Ela é uma ou mais palavras e/ou números e caracteres especiais que se refere ao tempo.

Exemplos de **Expressões Temporais (ETs)** são: “2018-05-18”, “ontem”, “manhã de ontem”, “12 horas atrás”, “Natal”, “Independência do Brasil”, “qid”, “bid”, “tid”, “três vezes por dia”, “12/12h”, “12/12 horas”, “mensalmente” e etc.

O processo de marcação temporal possui dois passos, a **extração** e a **normalização**.

A.3 Extração

A extração é o passo em que se identifica uma ET no texto. Por exemplo, na frase “O paciente não conseguiu dormir ontem”, a palavra “ontem” é uma **ET**, pois ela

se refere a um tempo.

A.4 Normalização

A normalização é o passo seguinte à extração. Nele, a ET é transformada para um padrão que pode ser reconhecido por um sistema computacional.

No nosso caso, usaremos a tag da ISO-TimeML “**TIMEX3**”. A tag TIMEX3 possui vários atributos. Entretanto, usaremos apenas alguns desses atributos, mais especificamente os atributos **type**, **value (val)**, **mod** e **freq**.

Uma expressão temporal marcada por uma tag TIMEX3 pode começar com **<TIMEX3>** e termina com **</TIMEX3>**. Entretanto, ela pode também seguir o padrão **<TIMEX3 atributos.../>**.

Na tabela abaixo são apresentados os atributos e os valores dos atributos que serão utilizados durante a anotação das expressões temporais.

Atributo	Quando usar	Valor de atributo
type	Em todas as tags TIMEX3	DATE, TIME, DURATION, SET
value	Em todas as tags TIMEX3	Depende da ET
mod	Quando necessário	NA, LESS, MORE, START, MID, END, APPROX, AFTER, BEFORE
freq	Apenas em conjunto com o atributo de valor SET	Valor inteiro e uma granularidade de tempo que represente a frequência com que a ET ocorre. Exemplo: “2x /dia”, deve ser marcado como “freq=2X”.

A.5 Orientações importantes sobre como fazer a normalização das expressões temporais

A.5.1 Idioma para normalização

É importante notar que a normalização sempre é feita utilizando a língua inglesa.

A.5.2 ET ruidosas

Quando encontrar ETs escritas fora do padrão (ruidosas), elas devem ser marcadas da forma ruidosa que foram escritas, seja esse ruído de qualquer tipo (sintático, semântico, de digitação etc.)

A.5.3 A justa medida de uma ET

Do ponto de vista teórico e prático, toda ET deve conter o máximo de informação importante para sua compreensão. Entretanto, ela deve ser também o mais definida possível. Vejamos os exemplos 1 e 2.

EX 1: “uso de ticlopidina **por 30 dias**, - retorno ao médico assistente.”

EX 2: “reavaliação com ginecologista **em 30 dias**.”

Os exemplos 1 e 2 mostram ETs completas e objetivas. No caso do EX 1, é importante marcar a preposição **por**, pois ela é quem indica a duração. No EX 2, a preposição **em** indica uma data futura em relação a data atual. Também é importante indicar a importância da palavra **dias**, pois ela é a unidade das ETs, uma vez que apenas o número 30 não representa uma ET. Os exemplos acima marcados na ferramenta MAE2¹ ficariam:

EX 1: <TIMEX3 id="T0" spans="8173~8184" text="por 30 dias" type="DURATION" val="P30D" mod="NA"/>

EX 2: <TIMEX3 id="T1" spans="18539~18549" text="em 30 dias" mod="NA" type="DATE" val="2017-09-25" />

No exemplo 2 acima, a data do sumário de alta é 25/08/2017.

A.6 Atributo type

Toda tag TIMEX3 deve obrigatoriamente conter o atributo type e seu valor de atributo. Estes podem ser de quatro tipo: **DATE**, **TIME**, **DURATION** e **SET** (SAURI et al., 2006; STYLER; WILLIAM et al., 2014).

DATE: Este tipo de ET define datas de calendário como “4 de janeiro” e outras expressões temporais que podem ser mapeadas para datas do calendário de forma concreta como “última semana”, “esse mês”, “próxima sexta-feira” ou “dessa vez”, ou em um sentido mais longo como “ultimamente”, ou “passado” (STYLER; WILLIAM et al., 2014).

Explicado de uma outra forma, o tipo DATE se refere a um período de tempo maior ou igual a um dia (STRÖTGEN, 2015), como por exemplo “10 dias”, “uma década”, “inverno”, “24 horas atrás” e etc.

- No que diz respeito a uma **data**, ela deve seguir o formato **YYYY-MM-DD**, onde **Y** equivale a **Ano**, **M** equivale a **Mês** e **D** equivale a **Dia**.
- Exemplos de data como 15 de maio de 18, 15/05/2018, 15/5/18, 15-05-2018, 15-5-18 etc. são todas normalizadas como **2018-05-15**.

¹ <http://keighrim.github.io/mae-annotation/>

- Uma data normalizada sempre segue o modelo ano com 4 dígitos, mês com 2 dígitos e dia com 2 dígitos. Sempre separado por hífen. Em um outro exemplo, a mesma data em diferentes formatos: 2/3/07, 02 03 2007, 02.03.07, 2.3.207, 2-03-2007 etc. todas essas formas devem ser normalizadas como **2007-03-02**.
- Se a ET possui apenas a informação do ano, por exemplo, deve seguir o formato **YYYY**

TIME: Este tipo de ET é usado para pontos de tempo específicos dentro de um dia. Exemplos: “três da tarde”, “23:45”, “20 minutos atrás”, “à tarde”, “almoço”, “café-da-manhã”, “o tempo da consulta” e etc (STYLER; WILLIAM et al., 2014).

Em outras palavras, uma expressão do tipo time refere-se a um ponto no tempo de qualquer granularidade menor que “dia” (STRÖTGEN, 2015).

- No caso de hora, minuto e segundo, estes devem seguir a mesma lógica de data, ou seja, **hh:mm:ss**. Da esquerda para a direita, deve vir a hora, depois o minuto e por fim o segundo, todos separados por dois pontos e com 2 dígitos. O formato utilizado deve ser o de 24 horas. Exemplos de ETs são: vinte e duas horas, 35 minutos e 8 segundos, 22:35:8, 02 e 35:8, 22h35m08s, 22 horas 35 min e 8 s, 22h35min 8 seg. etc. todas essas formas devem ser normalizadas como **22:35:08**
- Quando a informação possui dia e hora completa, se usa **2018-05-21T20:59:13**. Neste caso é necessário colocar um T maiúsculo entre o dia e a hora. Se a informação dos segundos não existir, ficará **2018-05-21T20:59**. Se a informação do minuto não existir, ficará **2018-05-21T20**

DURATION: Uma expressão de duração fornece informações sobre o comprimento de um intervalo. Eles podem se referir a intervalos de granularidades diferentes como, por exemplo: “três horas” ou “cinco anos” (STRÖTGEN, 2015).

Às vezes, você receberá uma única ET interpretada como refletindo um período de tempo, em vez de um ponto. São expressões temporais como, por exemplo: "por 24 horas" ou "Todo o mês de fevereiro".

- ET do tipo DURATION devem ser normalizadas seguindo o formato **P[n][Y/M/W/D]**, onde **P** significa período e o **n** é na verdade um número que indica a duração. A última parte do padrão indica a granularidade da duração **Y = ano**, **M = mês**, **W = semana** e **D = dia**. Os colchetes são apenas didáticos, não sendo usados na prática. Um exemplo seria “onze dias”, que é normalizado com val = ”P11D”, ou “dois meses e 3 dias”, que é normalizado para val = ”P2M3D”. A tabela 20 apresenta as abreviações de granularidade que devem ser usadas.

- A abreviação “M” pode significar Mês e Minuto, como na tabela 20. Essa diferenciação é mostrada no exemplo que segue. “**P20M**” = **20 meses**, entretanto “**PT20M**” = **20 minutos**. Assim, quando se referir a uma duração maior ou igual a 1 dia, deve-se iniciar a normalização com um “P” (Period), sendo que quando a duração for menor que 1 dia, deve-se iniciar a normalização com um “PT” (Period Time)

Table 20 – Abreviações

Abbreviation	Meaning
P	Período
Y	Ano
M	Mês
W	Semana
D	Dia
T	Tempo
H	Hora
M	Minuto
S	Segundo

SET: Esse tipo de ET é usado para cobrir expressões que fornecem um quantificador, um intervalo ou representam uma frequência como, por exemplo: "Três vezes por semana", "mensalmente" e "1/dia". Ele é diferente de DURATION ("a semana toda"), que só dá um intervalo (STYLER; WILLIAM et al., 2014). Apresentado de outra forma, uma expressão do tipo SET refere-se ao aspecto periódico de um evento, isto é, descreve um conjunto de horas ou datas (por exemplo, todas as segundas-feiras) ou uma frequência dentro de um intervalo de tempo (por exemplo, “duas vezes por semana”) (STRÖTGEN, 2015).

- A normalização das ET do tipo SET é feito através do atributo **freq**. Entretanto, vale lembrar que o atributo **freq** é usados apenas quando necessário, ou seja, é possível anotar uma ET do tipo SET sem o atributo **freq**.

A.7 Categorias de expressões temporais

Além dos 4 tipos de ETs utilizados no processo de normalização (DATE, TIME, DURATION, SET), há também 5 categorias de ETs no que diz respeito a sua completude de informação temporal (explícita, implícita, relativa, indeterminada e imprecisa). Vejamos cada uma delas.

ETs explícitas: Possui em si a informação completa para a normalização. Exemplos de ETs explícitas são: “2 de novembro de 2017”, “novembro de 2018”. Esses exemplos normalizados ficam 2017-11-02 e 2018-11, respectivamente.

ETs implícitas: Essa categoria foi projetada especificamente para datas nomeadas como o Natal, que sempre é comemorado dia 25 de dezembro. ETs implícitas também incluem, por exemplo, o Dia dos Pais, que sempre ocorre no 2º domingo de agosto. Considerando o ano de 2018, essas duas datas normalizadas ficam 2018-12-25 e 2018-08-12. No caso do Dia dos Pais, é sempre possível calcular sua data através de um cálculo de calendário, desde que se saiba o ano.

ETs relativas: Como pode ser visto nas duas categorias anteriores, elas não precisam de um contexto para serem normalizadas. Isso não acontece com ETs relativas e indeterminadas. Quando termos como “hoje”, “próximo mês” e “cinco dias depois” são encontrados, eles precisam de um tempo de referência para serem normalizados.

Por exemplo, no texto “Ele recebeu alta há dois dias, mas hoje voltou ao hospital”. Primeiro é necessário saber quando é “hoje”, e então, o termo “dois dias atrás” pode ser calculado em relação a “hoje”. Suponha que o tagger temporal (sistema que faz marcações temporais) seja capaz de identificar que “hoje” é “3 de outubro de 2016”, neste caso, “dois dias atrás” deve ser “1º de outubro de 2016”. Uma vez identificadas as expressões temporais, sua normalização seria 2016-10-03 e 2016-10-01, respectivamente.

ETs indeterminadas: Nesta categoria de ET, o contexto para a normalização também é importante. No entanto, essa categoria de ET agrupa expressões que possuem dados ausentes. Nestes casos, não é possível encontrar um tempo de referência no contexto como, por exemplo: “novembro” e “15 de novembro”. Assumindo que o contexto em que esses termos são inseridos não apresenta as informações do ano, a normalização dessas expressões seria XXXX-11 e XXXX-11-15, respectivamente.

- Caso a ET normalizada seja indeterminada, se deve substituir a parte indeterminada por **X** respeitando a quantidade de vezes que a informação completa exigiria. Por exemplo **XXXX-02-13**, **2018-05-21TXX:59:13**, **2018-XX-21T20:59:13**, **2018-05-XXT20:59** e **XXXX-XX-21**

ETs imprecisas: Existe ainda uma outra categoria de ET chamada imprecisa (TISSOT, 2016). Exemplos dessa categoria são “manter acompanhamento” e “apresentou melhora”, que são normalizadas como “FUTURE_REF” e “PAST_REF”, respectivamente. Uma ET imprecisa possui somente três valores possíveis, PAST_REF, PRESENT_REF e FUTURE_REF, onde: PAST_REF equivale a algo que aconteceu no passado; PRESENT_REF equivale a algo que aconteceu no presente e FUTURE_REF se refere a ao que acontecerá no futuro.

Sempre tomando uma data de referência para se referir a uma ET imprecisa. Por exemplo, se a data de referência é a data de alta de um paciente, a ET “manter acompanhamento” é normalizada com “FUTURE_REF”, pois o acompanhamento acontecerá depois da alta.

- No caso de ETs imprecisas, a normalização deverá ser feita utilizando os termos “PAST_REF” quando a ET imprecisa ocorreu no passado. “PRESENT_REF”, quando a ET imprecisa indicar a data presente. “FUTURE_REF”, quando a ET imprecisa indicar uma data futura. Uma ET imprecisa pode ser utilizada para indicar uma ET do tipo DATE, TIME, DURATION ou SET. É interessante tomar como ponto de referência para essa categoria de ET a data da evolução ou a data de alta de um sumário de alta. Na prática, se observará que raramente aparecerá uma ET imprecisa do tipo “PRESENT_REF”, pois relatos clínicos normalmente indicam um “tempo de referência” (a evolução indica o momento em que ela acontece, já o sumário de alta indica a data de alta do paciente)
- No que diz respeito à granularidade de uma ET, DATE também é usado para tipos muito genéricos de TIMEX3s, em que você pode não conseguir apontar para um dia, semana ou mês específico em um calendário, mas ainda pode ter uma ideia na timeline geral. Por exemplo, expressões como "no passado", "ultimamente", "no futuro" ou até "anteriormente" seriam TIMEX3s do tipo DATE (STYLER; WILLIAM et al., 2014). Colocado de forma diferente, as expressões temporais que fornecem detalhes minuto a minuto ou hora a hora são marcadas como TIME. O detalhe do dia-a-dia (ou maior) está marcado com DATE (STYLER; WILLIAM et al., 2014). Em todos esses casos, o tipo de ET é normalizado com um valor de ET imprecisa, como “PAST_REF, PRESENT_REF ou FUTURE_REF”.

A.8 Atributo value (val)

O atributo value também é obrigatório em todas as tags TIMEX3, assim como o atributo type.

A.9 Atributo mod

O atributo mod é usado apenas quando existe necessidade. Ele possui valores de atributo fixos, que são: “NA”, “LESS”, “MORE”, “START”, “MID”, “END”, “APPROX”, “BEFORE” e “AFTER”. A tabela 21 dá um exemplo de cada valor de atributo de mod:

Table 21 – Valores possíveis do atributo mod

Valor	Indicação	Exemplo	ET normalizada
NA	não se aplica	não se aplica	não se aplica
LESS		O tratamento durou quase 2 meses	<TIMEX3 type="DURATION" value="P2M" mod="LESS" text="quase 2 meses"/>
MORE		O tratamento durou mais de 2 meses	<TIMEX3 type="DURATION" value="P2M" mod="MORE" text="mais de 2 meses"/>
START		O tratamento ocorreu no começo de janeiro	<TIMEX3 type="DATE" value="XXXX-01" mod="START" text="começo de janeiro "/>
MID (Middle)		O tratamento ocorreu na metade de janeiro	<TIMEX3 type="DATE" value="XXXX-01" mod="MID" text="metade de janeiro "/>
END		O tratamento ocorreu no fim de janeiro	<TIMEX3 type="DATE" value="XXXX-01" mod="END" text="fim de janeiro "/>
APPROX	Quando o valor da ET indicar tempo aproximado de forma explícita no texto	O paciente se queixa da dor por quase uma semana	<TIMEX3 type="DURATION" value="P1W" mod="APPROX" text="quase uma semana "/>
AFTER		A crise ocorreu um pouco depois das 3 da manhã	<TIMEX3 type="TIME" value="03:00" mod="AFTER" text="depois das 3 da manhã "/>
BEFORE		A crise ocorreu um pouco antes das 3 da manhã	<TIMEX3 type="TIME" value="03:00" mod="BEFORE" text="antes das 3 da manhã "/>

A.10 O atributo freq

O atributo **freq** é usado apenas em conjunto como o type="SET", e apenas quando necessário. O atributo **freq** contém um valor inteiro e uma granularidade de tempo que representam a frequência com que a expressão temporal ocorre regularmente.

Exemplos:

```
<TIMEX3 type = "SET" value = "P1W" freq = "2X"> duas vezes por semana </
TIMEX3>
```

```
<TIMEX3 type = "SET" value = "P1D" freq = "2X"> 2x/dia </ TIMEX3>
```

A.11 O que não marcar?

Não se deve marcar um número sozinho.

1. Por exemplo: “**uso de ticlopidina por 30 dias, - retorno ao médico assistente**”. É errado marcar apenas o número, pois ele não indica necessariamente um ET. Neste caso, a marcação correta no MAE ficará.

```
<TIMEX3 id="T0" spans="8173 8184" text="por 30 dias" mod="NA" type="DURATION"
val="P30D" />
```

Seguem outros exemplos que evidenciam a importância do contexto na marcação.

2. uso de amitriptilina 75mg **por 1 mês**.

```
<TIMEX3 id="T2" spans="18819 18828" text="por 1 mês" mod="NA" val="P1M"
type="DURATION" />
```

3. uso de amitriptilina 75mg **há 1 mês**.

```
<TIMEX3 id="T3" spans="18857 18865" text="há 1 mês" type="DATE" val="2018-
05-11" mod="NA" />
```

4. uso de amitriptilina 75mg **em 1 mês**.

```
<TIMEX3 id="T4" spans="18894 18902" text="em 1 mês" type="DATE" val="2018-
07-11" mod="START" />
```

Há casos em que unidades de tempo não devem ser marcadas:

5. “história de hemorragia digestiva por up gástrica em vigência de aco **no mês de maio/03**.”. A marcação correta neste caso seria:

```
<TIMEX3 id="T5" spans="4985 4992" text="maio/03" type="DATE" mod="NA"
val="05-2003" />
```

No exemplo acima, as palavras **no mês** ou simplesmente **mês** não devem ser marcadas. Abaixo temos outro exemplo neste mesmo sentido.

6. pct vem a emergencia **dia 08.10** por angina de forte intensidade. <TIMEX3 id="T6" spans="5511 5516" text="**08.10**" mod="NA" type="DATE" val="2017-10-08" />

No exemplo acima, o **ano** é sabido pela data do sumário de alta, fornecido em sua versão completa.

Não marcar termos como

7. “retorno”, “segue em enfermaria”, “no momento”, “segue em <evento>”, “em curso”, “monitoração contínua”, “infusão contínua”, “contínuo”, “admissão” etc.
8. Não marcar ETs motivadas pelo tempo verbal como “apresentou”, “fará”, “manter” etc.

A.12 Protocolo de anotação

1. O anotador deve primeiramente fazer a **extração** de todas as ETs do relato clínico em que está trabalhando. Somente com todas as ETs extraídas (marcadas) é que se deve iniciar o processo de **normalização** do relato clínico. É importante que essa ordem seja respeitada, pois foi observado durante o piloto que esse processo gera melhores resultados de normalização. Isso acontece por haver casos em que as ETs iniciais do relato clínico não possuem informações completas para a normalização. Entretanto, ETs do final do texto possuem as informações completas que podem ser usadas nas ETs do início do relato.
2. É importante sempre buscar fazer a normalização com o máximo de informações possíveis. Se o texto clínico possuir informação de ano, mês e dia, assim como hora e minuto, eles devem fazer parte da normalização da ET. Caso a informação completa não exista, neste caso, se normaliza a ET apenas com a informação disponível.
3. Deve-se utilizar a normalização de ETs do tipo imprecisas “PAST_REF”, “FUTURE_REF” e “PRESENT_REF” apenas quando não for possível fazer a normalização de uma maneira mais específica. Em outras palavras, esse tipo de normalização deve ser usado apenas quando outras opções de normalização já tiverem sido esgotadas, ou seja, não existirem.
4. Sempre se deve pegar a data mais recente como a data de referência, principalmente em sumários de alta. Por exemplo no texto abaixo, a data mais recente é **15/01/2015**, ela deve se usada como referência para calcular as ETs **há 20 anos**, que normalizada fica type=”DATE” val=”1995”.

“pai faleceu por ca de próstata, mãe com alzheimer, também falecida. chv : ex tabagista (60 maços / ano, parou **há 20 anos**), ex etilista **há 20 anos**. refere alimentação saudável (2 refeições por dia - café da manhã e janta). realiza caminhada todos os dias (mínimo 30 minutos). rs: sem particularidades. nega queixas urinárias e gastrointestinais. exames laboratoriais :**15/01/2015** cr 0.8 triglicerideos 74 transaminase piruvica 24 hdl 48 sodio 137 potassio 5.1 transaminase oxalacetica 27 colesterol total 123 ldl 60 glicose 80 a1c 6.49% oftalmo = 08 /2014 = tudo normal (angiografia) exame do pé = 12/2014 = normal 12/11/2014 = relação albumina / creatinina 54.32 mg/g o # ef beg, normocorado, hidratado, anictérico, acianótico, afebril.”

5. Apesar de inicialmente normalizações como **1996-XX-XX** e **XXXX-XX-XXT19:00**, terem sido estimuladas, após uma reavaliação, elas foram descartadas, pois não trazem informação adicional. Assim, seu formato correto deve ser **1999** e **19:00**.

Via de regra, quando a parte não conhecida/indeterminada da ET for à direita da parte conhecida, não se deve completá-la com “X”. Por exemplo, 2018-XX-XX e 2018-07-XX. Estes exemplos não devem ter sua parte indeterminada completada com “X”. Sua correta normalização deve ser **2018** e **2018-07** respectivamente.

Por outro lado, se a parte indeterminada da ET for à esquerda da parte conhecida da ET, ela deve ser completada com ”X”. Exemplos são: 12-31 e 29 devem ser normalizados como **XXXX-12-31** e **XXXX-XX-29**. Um outro exemplo com contexto seria “no terceiro dia do mês de maio o procedimento foi realizado.”. A data do último exemplo deve ser normalizada como **XXXX-05-03**.

6. Tomar cuidado com frases como “insulina nph 20u à noite”. Elas podem parecer se tratar de uma ET do tipo **TIME**, mas na verdade são uma frequência, do tipo **SET**.
7. As possibilidades do atributo “**mod**” deve ser usados apenas quando houver uma palavra ou contexto que indique a necessidade de sua aplicação. Por exemplo, “**APPROX**” deve ser usado apenas quando houver uma palavra no texto que expresse incerteza. Se o anotador tiver a impressão ou entender que deve usá-lo, mas não há uma palavra ou conjunto de palavras que não expressam essa incerteza, então, não se deve usar o “**APPROX**”. Exemplos de palavras que indicam a necessidade de usar o “**APPROX**” são: aproximadamente, há cerca de, por volta de, há 6-7 anos etc.
8. Não se deve relacionar a(s) mesma(s) palavra(s) em duas ETs diferentes, como em “**08 meses**” e “**há cerca de 08**”, onde “**08**” está presente nas duas ETs. Segue trecho do relato: “parou de usar **há cerca de 08 meses** devido a mal estar durante

a madrugada (hipoglicemia)”. Neste exemplo o **negrito** indica uma ET marcada e o sublinhado indica a outra ET marcada.

9. Os casos de “**60 maços/ano**”, “**41 anos/maço**” etc., não devem ser marcados como SET, pois são equivalentes aos casos que indicam a quantidade de batimentos por minuto (bpm), por exemplo, que aparentam indicar tempo, mas na verdade indica a carga tabágica do paciente, semelhante a uma unidade (bpm). Foi acordado que esse tipo de informação não será marcado.
10. ETs que possuem “há cerca de” (por volta de/em torno de <número> <unidade de tempo>) podem significar dois tipos diferentes de ETs: DATE ou DURATION. Por exemplo:

“realiza tratamento com insulina **há cerca de** 03 anos, parou de usar há cerca de 08 meses devido a mal estar”

No caso destacado em **negrito**, se trata de uma ET do tipo DURATION. Já no caso destacado em sublinhado, se trata de uma ET do tipo DATE. É um caso ambíguo. Neste caso, ficou decidido que se deve seguir o contexto do relato clínico. Se o contexto indica um ponto no tempo, a ET deve ser marcada com DATE. Entretanto, se o contexto indicar que a ET significa uma duração, ela deve ser marcada como DURATION. Tomou-se essa decisão com base na relevância da informação temporal para os profissionais de saúde.

11. Os casos “**à noite**”, “**de manhã**”, “**pela manhã**”, “**à tarde**”, “**no almoço**” etc., podem ser marcados como do tipo **DATE** ou **SET**, dependendo do contexto.

Por exemplo, “**à noite**”, pode ser do tipo **DATE**, se a prescrição de um medicamento for para ser tomado apenas uma vez. Entretanto, se o contexto do relato clínico indicar que o medicamento deve ser tomado todas as noites, então, essa ET deverá ser marcado como tipo **SET**.

The guideline was created during the whole annotation process, especially during the pilot tests. All decisions were based on other guidelines studied by the adjudicator, on temporal information needed by the health professionals (annotators) in their day-to-day duties and in consensus with the annotators and the professor who was coordinating the annotation process. The annotation protocol is the part that illustrates best the adjustments needed appointed by the annotators according to their experience.

APPENDIX B – Guideline

The original guideline was written in Brazilian Portuguese. This appendix is a free translation of the original.

B.1 Motivation

The temporal tagging process is important for computational systems whose input is text. Among these systems we can list the ones destined for question / answering, chat bots, timeline creation and summarization.

For this work we assume the fundamental assumption that: “If a human can determine a value for the temporal expression, it should be tagged” (FERRO et al., 2001). In other words, if a person understands that a word or set of words denotes time, that temporal expression must be marked. However, it is necessary that this marking also allows its normalization. In this sense, it is also important to point out that the temporal expressions that were considered useful by health specialists prevailed. These professionals participated in the construction of this guideline, just as they also used it.

B.2 Temporal expression

The basic unit for the temporal tagging process is the Temporal Expression (TE). It is one or more words and/or numbers and special characters that refers to time.

Examples of Temporal Expressions are: "2018-05-18", "yesterday", "yesterday morning", "12 hours ago", "Christmas", "Independence of Brazil", "qid", "bid", "tid", "three times a day", "12/12h", "12/12 hours", "monthly", and so on.

The temporal tagging process has two steps, extraction and normalization.

B.3 Extraction

Extraction is the step when a TE is identified in the text. For example, in the phrase "The patient could not sleep yesterday", the word "yesterday" is a TE, since it refers to time.

B.4 Normalization

Normalization is the next step to extraction. In it, a TE is transformed to a pattern that can be recognized by a computational system.

In our case, we will use a ISO-TimeML tag called "**TIMEX3**". The **TIMEX3** tag has several attributes. However, we will only use some of these attributes, specifically the **type**, **value** (val), **mod**, and **freq** attributes.

A temporal expression marked by a TIMEX3 tag can start with `<TIMEX3>` and ends with `</TIMEX3>`. However, it can also follow the standard `<TIMEX3 attributes ... />`

The table below shows the attributes and values of the attributes that will be used during the annotation of temporal expressions.

Attribute	When to use	Attribute value
type	In all tags TIMEX3	DATE, TIME, DURATION, SET
value	In all tags TIMEX3	Depends on the TE
mod	When necessary	NA, LESS, MORE, START, MID, END, APPROX, AFTER, BEFORE
freq	Only in conjunction with the SET value attribute	Integer value and a granularity of time that represents the frequency which the TE occurs. Example: "2x /day should be marked as "freq=2X".

B.5 Important guidelines on how to normalize temporal expressions

B.5.1 Language for normalization

It is important to note that normalization is always done using the English language.

B.5.2 Noisy temporal expression

When encountering TEs written in a non-standard way (noisy), it should not be corrected, it must be tagged keeping the non-standard way (noisy), be it of any type (syntax, semantics, typing, etc.)

B.5.3 The fair measure of a TE

From a theoretical and practical point of view, every TE must contain as much information as is important for its understanding. However, it must be as objective as

possible. Let's look at examples 1 and 2.

EX 1: "use of ticlopidine **for 30 days**, - return to attending physician." **EX 2:** "re-evaluation with gynecologist **in 30 days**."

Examples 1 and 2 show complete and objective TEs. In the case of EX 1, it is important to tag the preposition **for**, since it is the one that indicates the duration. In EX 2, the preposition **in** indicates a future date in relation to the current date. It is also important to indicate the importance of the word **days**, since it is the unit of the TEs, since only the number 30 does not represent a TE. The examples above tagged with the MAE tool would be:

EX 1: <TIMEX3 id="T0" spans="8173~8184" text="**for 30 days**" type="DURATION" val="P30D" mod="NA"/>

EX 2: <TIMEX3 id="T1" spans="18539~18549" text="**in 30 days**" mod="NA" type="DATE" val="2017-09-25" />

In example 2 above, the date of the discharge summary is 08/25/2017.

B.6 The type attribute

Every TIMEX3 tag must contain the type attribute and its attribute value. These can be of four types: **DATE**, **TIME**, **DURATION** and **SET** (SAURI et al., 2006; STYLER; WILLIAM et al., 2014).

DATE: This type of TE defines calendar date as "last week," "this month," "next Friday," or "this time," in a longer sense like "lately" or "past" (STYLER; WILLIAM et al., 2014).

Explained in a different way, the DATE type refers to a period of time greater than or equal to a day (STRÖTGEN, 2015), such as "10 days", "a decade", "winter", "24 hours ago", etc.

- With respect to a date, it must follow the format **YYYY-MM-DD**, where **Y** is **Year**, **M** is **Month**, and **D** is **Day**.
- Examples of date such as May 15, 18, 15/05/2018, 15/5/18, 15-05-2018, 15-5-18 etc. are all normalized as **2018-05-15**.
- A normalized date always follows the year model with 4 digits, month with 2 digits and day with 2 digits. Always separated by hyphen. In another example, the same date in different formats: 2/3/07, 02 03 2007, 02.03.07, 2.3.207, 2-03-2007 etc. all such forms should be normalized as **2007-03-02**.

- If the TE has only the information of the year, for example, it must follow the format **YYYY**.

TIME: This type of TE is used for specific time points within a day. Examples: "three in the afternoon", "23:45", "20 minutes ago", "afternoon", "lunch", "breakfast", "consultation time" and etc (STYLER; WILLIAM et al., 2014).

In other words, a time-type temporal expression refers to a point in time of any granularity less than "day" (STRÖTGEN, 2015).

- In the case of hour, minute and second, they must follow the same date logic, that is, **hh:mm:ss**. From left to right, the time must come, then the minute and finally the second, all separated by a colon and with 2 digits. The format used should be 24 hours. Examples of TEs are: twenty-two hours, 35 minutes and 8 seconds, 22:35:8, 02 and 35:8, 22h35m08s, 22 hours 35 mins and 8s, 22h35min 8s. etc. all these forms must be normalized as **22:35:08**
- When the information has full day and time, use **2018-05-21T20:59:13**. In this case it is necessary to place a capital T between day and time. If the information of the seconds does not exist, it will be **2018-05-21T20:59**. If the minute information does not exist, it will be **2018-05-21T20**

DURATION: A duration expression provides information about the length of a range. They can refer to intervals of different granularities such as "three hours" or "five years" (STRÖTGEN, 2015).

Sometimes you will get a single TE interpreted as reflecting a period of time rather than a period. They are temporal expressions such as "for 24 hours" or "All the month of February".

- Temporal expressions of type DURATION should be normalized following the format **P[n][Y/M/W/D]**, where **P** means **period** and **n** is actually a number indicating the duration. The last part of the pattern indicates the granularity of the duration **Y = year**, **M = month**, **W = week** and **D = day**. The brackets are just didactic, not being used in practice. An example would be "eleven days", which is normalized with val="P11D", or "two months and 3 days", which is normalized to val="P2M3D". The table 22 shows the granularity abbreviations that should be used.
- The abbreviation "M" can mean Month and Minute, as in table 22. This differentiation is shown in the following example. **"P20M" = 20 months**, meanwhile **"PT20M" = 20 minutes**. Thus, when referring to a duration greater than or equal

to 1 day, the normalization should begin with a "P" (Period), and when the duration is less than 1 day, normalization with a "PT"(Period Time)

Table 22 – Abbreviation table

Abbreviation	Meaning
P	Period
Y	Year
M	Month
W	Week
D	Day
T	Time
H	Hour
M	Minute
S	Second

SET: This type of TE is used to cover expressions that provide a quantifier, a range, or represent a frequency such as "Three times a week", "monthly" and "1/day". It is different from DURATION ("all week"), which only gives an interval (STYLER; WILLIAM et al., 2014). Presented in another way, an expression of the SET type refers to the periodic aspect of an event, that is, it describes a set of times or dates (for example, every Monday) or a frequency within a time interval (for example, "twice a week") (STRÖTGEN, 2015).

- The normalization of SET type TEs is done through the **freq** attribute. However, it is worth remembering that the freq attribute is used only when necessary, it means, it is possible to write a SET-type TE without the **freq** attribute.

B.7 Categories of temporal expressions

In addition to the four types of TEs used in the normalization process (DATE, TIME, DURATION and SET), there are also five categories of TEs with respect to their temporal information completeness (explicit, implicit, relative, indeterminate and imprecise). Let's look at each one of them.

Explicit TEs: It has in itself the complete information for normalization. Examples of explicit TEs are: "November 2, 2017", "November 2018". These normalized examples are 2017-11-02 and 2018-11, respectively.

Implicit TEs: This category was designed specifically for dates named as Christmas, which is always celebrated on December 25th. Implicit TEs also include, for example, Father's Day (in Brazil), which always occurs on the 2nd Sunday of August. Considering the year 2018, these two normalized dates are 2018-12-25 and 2018-08-12. In the case

of Father's Day, it is always possible to calculate your date by a calendar calculation, provided the year is known.

Relative TEs: As can be seen in the two previous categories, they do not need a context to be normalized. This does not happen with relative and indeterminate TEs. When terms like "today", "next month" and "five days later" are found, they need a reference time to be normalized.

For example, in the text "He was discharged two days ago, but today he returned to the hospital." First it is necessary to know when it is "today", and then, the term "two days ago" can be calculated in relation to "today". Suppose the temporal tagger is able to identify that "today" is "October 3, 2016," in which case "two days ago" should be "October 1, 2016." Once the temporal expressions were identified, their normalization would be 2016-10-03 and 2016-10-01, respectively.

Undetermined TEs: In this TE category, the context for normalization is also important. However, this TE category groups expressions that have missing data. In these cases, it is not possible to find a reference time in the context such as "November" and "November 15". Assuming that the context in which these terms are inserted does not display the year information, the normalization of these expressions would be XXXX-11 and XXXX-11-15, respectively.

- If the normalized TE is indeterminate, the undetermined part must be replaced by **X** respecting the amount of times the complete information would require. For example: **XXXX-02-13**, **2018-05-21TXX:59:13**, **2018-XX-21T20:59:13**, **2018-05-XXT20:59** and **XXXX-XX-21**

Imprecise TEs: There is yet another category of TE called imprecise (TISSOT, 2016). Examples of this category are "in 6 months", "today" and "recent", which are normalized to "FUTURE_REF", "PRESENT_REF" and 'PAST_REF', respectively. An imprecise TE has only three possible values, PAST_REF, PRESENT_REF, and FUTURE_REF, where: PAST_REF equals something that happened in the past; PRESENT_REF equals something that happened in the present, and FUTURE_REF refers to what will happen in the future.

It is always important to take a reference date to refer to an imprecise TE. For example, if the reference date is the return of a patient to see the physician, the TE "in 6 months" is normalized to "FUTURE_REF" because follow-up will occur after discharge.

- In the case of imprecise TEs, normalization should be done using the terms "PAST_REF" when the imprecise TE has occurred in the past. "PRESENT_REF", when the imprecise TE indicates the present date. "FUTURE_REF", when the imprecise

TE indicates a future date. An imprecise TE can be used to indicate a DATE, TIME, DURATION or SET type TE. It is interesting to take as reference point for this TE category the evolution date or the discharge date of a discharge summary. In practice, it will be noted that an imprecise ET of the "PRESENT_REF" type will rarely appear, as clinical reports usually indicate a "reference time" (evolution indicates when it happens, while the discharge summary indicates the discharge date of the patient)

- Regarding the granularity of a TE, DATE is also used for very generic types of TIMEX3s, where you may not be able to point to a specific day, week, or month on a calendar, but may still have an idea in the overall timeline. For example, expressions like "in the past", "lately", "in the future" or even "before" would be TIMEX3s of type DATE (STYLER; WILLIAM et al., 2014). Placed differently, timestamps that provide minute-by-minute or hour-by-hour details are marked as TIME. The detail of day-to-day (or greater) is marked with DATE (STYLER; WILLIAM et al., 2014). In all these cases, the TE type is normalized with an imprecise TE value, such as "PAST_REF, PRESENT_REF or FUTURE_REF".

B.8 The value (val) attribute

The value attribute is also required in all TIMEX3 tags, as well as the type attribute.

B.9 The mod attribute

The mod attribute is used only when there is a need. It has fixed attribute values, which are: "NA", "LESS", "MORE", "START", "MID", "END", "APPROX", "BEFORE", and "AFTER". Examples are presented in table 23.

Table 23 – Mod options

Mod value	Note	Example	Normalized TE
NA	not applicable	not applicable	not applicable
LESS		The treatment lasted almost 2 months	<TIMEX3 type="DURATION" value="P2M" mod="LESS" text="almost 2 months"/>
MORE		The treatment lasted more than 2 months	<TIMEX3 type="DURATION" value="P2M" mod="MORE" text="more than 2 months"/>
START		The treatment occurred in the beginning of January	<TIMEX3 type="DATE" value="XXXX-01" mod="START" text="beginning of January "/>
MID (Middle)		The treatment took place in the middle of January	<TIMEX3 type="DATE" value="XXXX-01" mod="MID" text="middle of January "/>
END		The treatment took place at the end of January	<TIMEX3 type="DATE" value="XXXX-01" mod="END" text="end of January "/>
APPROX	When the TE value indicates explicit approximate time in text	The patient complains of pain for almost a week	<TIMEX3 type="DURATION" value="P1W" mod="APPROX" text="almost a week "/>
AFTER		The crisis took place a little bit after 3 AM	<TIMEX3 type="TIME" value="03:00" mod="AFTER" text="after 3 AM "/>
BEFORE		The crisis took place a little bit before 3 AM	<TIMEX3 type="TIME" value="03:00" mod="BEFORE" text="before 3 AM "/>

B.10 The freq attribute

The freq attribute is only used with the type = "SET", and only when necessary. The freq attribute contains an integer value and a granularity of time that represents the frequency with which the temporal expression occurs regularly.

Examples:

```
<TIMEX3 type = "SET" value = "P1W" freq = "2X"> twice a week </ TIMEX3>
```

```
<TIMEX3 type = "SET" value = "P1D" freq = "2X"> 2x/day </ TIMEX3>
```

B.11 What not to tag?

You should not tag a number alone.

1. For example: "use of ticlopidine for 30 days, - return to the attending physician". It is wrong to just tag the number because it does not necessarily indicate a TE. In this case, the correct tag in the MAE tool will remain.

```
<TIMEX3 id="T0" spans="8173 8184" text="for 30 days" mod="NA" type="DURATION" val="P30D" />.
```

Following are other examples that demonstrate the importance of context in tagging.

2. use of amitriptyline 75mg for 1 month.

```
<TIMEX3 id="T2" spans="18819 18828" text="for 1 month" mod="NA" val="P1M" type="DURATION" />
```

3. use of amitriptyline 75mg 1 month ago.

```
<TIMEX3 id="T3" spans="18857 18865" text="1 month ago" type="DATE" val="2018-05-11" mod="NA" />
```

4. use of amitriptyline 75mg in 1 month.

```
<TIMEX3 id="T4" spans="18894 18902" text="in 1 month" type="DATE" val="2018-07-11" mod="START" />
```

There are cases where time units should not be tagged:

5. "History of digestive hemorrhage by gastric upset during **the month of May/03**." The correct marking in this case would be:

```
<TIMEX3 id="T5" spans="4985 4992" text="May/03" type="DATE" mod="NA" val="2003-05" />
```

In the example above, the words "the month of" or simply "month" should not be tagged. Below we have another example in this sense.

6. pct (patient) comes the emergency **day 08.10** by angina of strong intensity.

```
<TIMEX3 id="T6" spans="5511 5516" text="08.10" mod="NA" type="DATE" val="2017-10-08" />
```

In the above example, the year is known by the discharge summary date, provided in its full version.

Do not tag terms as

7. "return", "follow in ward", "at the moment", "follows in <event>", "ongoing", "continuous monitoring", "continuous infusion", "continuous", "admission", etc.
8. Do not tag TEs motivated by verbal tense as "presented", "will", "maintain", and so on.

B.12 Annotation protocol

1. The annotator must first extract all TEs from the clinical report he is working on. Only with all TEs extracted (marked) should the process of normalization of the clinical report begin. It is important that this order be respected, as it was observed during the pilot tests that this process generates better normalization results. This is because there are cases in which the initial TEs of the clinical report do not have complete information for normalization (specially the relative TEs). However, there are cases where TEs at the end of the text have the complete information that can be used in TEs at the beginning of the clinical report.
2. It is important to always seek to normalize with as much information as possible. If the clinical text contains year, month and day information, as well as hour and minute, they should be part of TE normalization. If the complete information does not exist, in this case, the TE is normalized only with the available information.
3. Normalization of imprecise type TEs "PAST_REF", "FUTURE_REF" and "PRESENT_REF" should be used only when normalization can not be done in a more specific way. In other words, this type of normalization should only be used when other normalization options have already been exhausted, that is, they do not exist.
4. One should always pick the most recent date as the reference date, especially in discharge summaries. For example in the text below, the most recent date is "01/15/2015", it should be used as a reference to calculate the TEs "20 years ago", which normalized becomes type="DATE" val="1995".

Example in Portuguese.

“pai faleceu por ca de próstata, mãe com alzheimer, também falecida. chv : ex tabagista (60 maços / ano, parou há 20 anos), ex etilista há 20 anos. refere alimentação

saudável (2 refeições por dia - café da manhã e janta). realiza caminhada todos os dias (mínimo 30 minutos). rs: sem particularidades. nega queixas urinárias e gastrointestinais. exames laboratoriais :15/01/2015 cr 0.8 triglicerideos 74 transaminase piruvica 24 hdl 48 sodio 137 potassio 5.1 transaminase oxalacetica 27 colesterol total 123 ldl 60 glicose 80 a1c 6.49% oftalmo = 08 /2014 = tudo normal (angiografia) exame do pé = 12/2014 = normal 12/11/2014 = relação albumina / creatinina 54.32 mg/g o # ef beg, normocorado, hidratado, anictérico, acianótico, afebril.”

5. Although initially normalizations like 1996-XX-XX and XXXX-XX-XXT19: 00 were stimulated, after a reevaluation, they were discarded because they do not bring additional information. Thus, its correct format should be 1999 and 19:00.

As a rule, when the unknown/undetermined TE part is to the right of the known part, it should not be completed by "X". For example, 2018-XX-XX and 2018-07-XX. These examples should not have their undetermined part completed with "X". Its correct normalization should be 2018 and 2018-07 respectively.

On the other hand, if the undetermined part of the TE is to the left of the known part of the TE, it must be completed with "X". Examples are: 12-31 and 29 should be normalized as XXXX-12-31 and XXXX-XX-29. Another example with context would be "on the third day of May the procedure was performed." The date of the last example should be normalized to XXXX-05-03.

6. Beware of phrases like "insulin nph 20u at night". They may look like a TIME type TE, but they are actually a SET type frequency.
7. The possibilities of the "mod" attribute should be used only when there is a word or context that indicates the need for its application. For example, "APPROX" should only be used when there is a word in the text that expresses uncertainty. If the annotator has the impression or understands that he should use it, but there is no word or set of words that do not express this uncertainty, then one should not use APPROX. Examples of words that indicate the need to use "APPROX" are: approximately, about, around, for 6-7 years etc.
8. You should not list the same word (s) in two different TEs, as in "**08 months**" and "**about 08**", where "**08**" is present in the two TEs. The following is an excerpt from the report: "stopped using it for **about 08 months** due to malaise during the night (hypoglycemia)". In this example the bold indicates a tagged TE and the underline indicates the other TE tagged.
9. The cases of "60 packs/year" (60 maços/ano), "41 years/pack" (40 maços/ano) etc., should not be marked as TE, as they are equivalent to the cases that indicate the number of beats per minute (bpm), for example, which appear to indicate time,

but actually indicates the patient's smoking load, similar to one unit (bpm). It was agreed that this type of information will not be tagged.

10. TEs that have "há cerca de" (around/about <number> <time unit>) can mean two different types of TEs: DATE or DURATION. For example:

(realiza tratamento com insulina **há cerca** de 03 anos, parou de usar há cerca de 08 meses devido a mal estar)

"Performed insulin treatment for **about** 03 years, stopped using it for about 08 months due to malaise"

In the case highlighted in bold, this is a DURATION type TE. In the case highlighted in underline, this is a DATE type TE. This is an ambiguous case. In this case, it was decided that the context of the clinical report should be followed. If the context indicates a point in time, the TE must be marked with DATE. However, if the context indicates that TE means a duration, it must be tagged as DURATION. This decision was made based on the relevance of time information to health professionals.

11. "Evening", "morning", "afternoon", "lunch", etc. can be tagged as DATE or SET, depending on the context.

For example. "at night" may be of the DATE type if the prescription of a drug is to be taken only once. However, if the context of the clinical report indicates that the medication should be taken every night, then this TE should be tagged as SET type.

The guideline was created during the whole annotation process, especially during the pilot tests. All decisions were based on other guidelines studied by the adjudicator, on temporal information needed by the health professionals (annotators) in their day-to-day duties and in consensus with the annotators and the professor who was coordinating the annotation process. The annotation protocol is the part that illustrates best the adjustments needed appointed by the annotators according to their experience.

Bibliography

- AGIRRE, E.; MÀRQUEZ, L.; WICENTOWSKI, R. Proceedings of the fourth international workshop on semantic evaluations (semeval-2007). In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. [S.l.: s.n.], 2007. Cited on page 43.
- ALLEN, J. F. Maintaining knowledge about temporal intervals. *Communications of the ACM*, ACM, v. 26, n. 11, p. 832–843, 1983. Cited on page 23.
- ALONSO, O. et al. Temporal information retrieval: Challenges and opportunities. *Twaw*, v. 11, p. 1–8, 2011. Quoted 2 times on pages 14 and 23.
- ALTUNA, B.; ARANZABE, M. J.; ILARRAZA, A. D. de. Eusheidelttime: Time expression extraction and normalisation for basque. *Procesamiento del Lenguaje Natural*, v. 59, p. 15–22, 2017. Cited on page 16.
- AZEVEDO, R. F. de et al. Temporal tagging of noisy clinical texts in brazilian portuguese. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2018. p. 231–241. Quoted 3 times on pages 52, 73, and 118.
- BAPTISTA, J. Evaluation of finite-state lexical transducers of temporal adverbs for lexical analysis of portuguese texts. In: SPRINGER. *International Workshop on Computational Processing of the Portuguese Language*. [S.l.], 2003. p. 235–242. Cited on page 17.
- BAPTISTA, J. Some families of compound temporal adverbs in portuguese. In: *Proceedings of Workshop on Finite-State Methods for Natural Language Processing*. [S.l.: s.n.], 2003. p. 97–104. Cited on page 17.
- BETHARD, S. et al. Semeval-2015 task 6: Clinical tempeval. In: *SemEval@ NAACL-HLT*. [S.l.: s.n.], 2015. p. 806–814. Quoted 2 times on pages 44 and 48.
- BETHARD, S. et al. Semeval-2016 task 12: Clinical tempeval. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. [S.l.: s.n.], 2016. p. 1052–1062. Cited on page 48.
- BETHARD, S. et al. Semeval-2017 task 12: Clinical tempeval. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. [S.l.]: Association for Computational Linguistics, 2017. p. 565–572. Cited on page 49.
- CAMPILLOS, L. et al. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, Springer, v. 52, n. 2, p. 571–601, 2018. Cited on page 44.
- CAMPOS, R. et al. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, ACM, v. 47, n. 2, p. 15, 2015. Quoted 2 times on pages 14 and 16.

- CHANG, A. X.; MANNING, C. D. SUTIME: A library for recognizing and normalizing time expressions. In: *LREC*. [S.l.: s.n.], 2012. v. 2012, p. 3735–3740. Cited on page 45.
- COHAN, A.; MEURER, K.; GOHARIAN, N. Guir at semeval-2016 task 12: Temporal information processing for clinical narratives. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. [S.l.: s.n.], 2016. p. 1248–1255. Cited on page 48.
- COSTA, F.; BRANCO, A. Timebankpt: A timeml annotated corpus of portuguese. In: *LREC*. [S.l.: s.n.], 2012. p. 3727–3734. Cited on page 44.
- CRAVEIRO, O.; MACEDO, J.; MADEIRA, H. Portexto: sistema de anotação/extracção de expressões temporais. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca*, 2008. Cited on page 17.
- CRAVEIRO, O.; MACEDO, J.; MADEIRA, H. Leveraging temporal expressions for segmented-based information retrieval. In: IEEE. *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*. [S.l.], 2010. p. 754–759. Cited on page 17.
- CRAVEIRO, O.; MACEDO, J.; MADEIRA, H. It is the time for portuguese texts! In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2012. p. 106–112. Quoted 3 times on pages 16, 17, and 18.
- DUNN, O. J. Multiple comparisons using rank sums. *Technometrics*, Taylor & Francis Group, v. 6, n. 3, p. 241–252, 1964. Cited on page 86.
- FERREIRA, L.; TEIXEIRA, A.; CUNHA, J. P. da S. Information extraction from portuguese hospital discharge letters. *Evolution*, v. 8, n. 998, p. 506, 2010. Cited on page 18.
- FERRO, L. et al. Tides 2005 standard for the annotation of temporal expressions. 2005. Quoted 2 times on pages 9 and 28.
- FERRO, L. et al. Tides temporal annotation guidelines-version 1.0. 2. *The MITRE Corporation, McLean-VG-USA*, 2001. Quoted 6 times on pages 20, 27, 28, 59, 120, and 132.
- GERTZ, J. S. J. Z. M. Heideltime: Tuning english and developing spanish resources for tempeval-3. *Atlanta, Georgia, USA*, p. 15, 2013. Cited on page 45.
- GROUIN, C.; MORICEAU, V. Limsi at semeval-2016 task 12: machine-learning and temporal information to identify clinical events and time expressions. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. [S.l.: s.n.], 2016. p. 1225–1230. Cited on page 48.
- GROUP, T. W. et al. *Guidelines for temporal expression annotation for english for tempeval 2010*. 2009. Quoted 4 times on pages 9, 28, 30, and 59.
- GUPTA, N.; JOSHI, A.; BHATTACHARYYA, P. A temporal expression recognition system for medical documents by taking help of news domain corpora. In: *12th International Conference on Natural Language Processing (ICON)*. [S.l.: s.n.], 2015. Quoted 2 times on pages 46 and 118.

HAGÈGE, C.; BAPTISTA, J.; MAMEDE, N. Identificação, classificação e normalização de expressões temporais do português: A experiência do segundo harem eo futuro. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguatca*, p. 33–54, 2008. Quoted 3 times on pages 92, 112, and 113.

HAGEGE, C.; BAPTISTA, J.; MAMEDE, N. Portuguese temporal expressions recognition: from te characterization to an effective ter module implementation. In: IEEE. *Information and Human Language Technology (STIL), 2009 Seventh Brazilian Symposium in*. [S.l.], 2009. p. 36–43. Quoted 2 times on pages 16 and 18.

HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009. Cited on page 44.

HAMON, T.; GRABAR, N. Tuning heideltime for identifying time expressions in clinical texts in english and french. *EACL 2014*, p. 101–105, 2014. Quoted 3 times on pages 45, 52, and 118.

HAO, T. et al. A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts. *BMC medical informatics and decision making*, BioMed Central, v. 18, n. 1, p. 22, 2018. Cited on page 46.

HAO, T.; RUSANOV, A.; WENG, C. Extracting and normalizing temporal expressions in clinical data requests from researchers. In: SPRINGER. *International Conference on Smart Health*. [S.l.], 2013. p. 41–51. Cited on page 46.

IV, W. F. S. et al. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 2, p. 143–154, 2014. Quoted 2 times on pages 44 and 48.

KOCOŃ, J.; MARCIŃCZUK, M. Supervised approach to recognise polish temporal expressions and rule-based interpretation of timexes. *Natural Language Engineering*, Cambridge University Press, v. 23, n. 3, p. 385–418, 2017. Cited on page 16.

KOLOMIYETS, O.; MOENS, M.-F. Kul: recognition and normalization of temporal expressions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 5th International Workshop on Semantic Evaluation*. [S.l.], 2010. p. 325–328. Cited on page 14.

KREIMEYER, K. et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, Elsevier, v. 73, p. 14–29, 2017. Quoted 3 times on pages 18, 42, and 43.

LEE, H.-J. et al. Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. [S.l.: s.n.], 2016. p. 1292–1297. Cited on page 49.

LEE, K. et al. Context-dependent semantic parsing for time expressions. In: *ACL (1)*. [S.l.: s.n.], 2014. p. 1437–1447. Cited on page 45.

- LEEUWENBERG, A.; MOENS, M.-F. Kuleuven-liir at semeval-2017 task 12: Cross-domain temporal information extraction from clinical records. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. [S.l.: s.n.], 2017. p. 1030–1034. Cited on page 49.
- LI, H. et al. Chinese temporal tagging with heidelttime. In: *EACL*. [S.l.: s.n.], 2014. v. 2014, p. 133–7. Cited on page 45.
- LI, M.; PATRICK, J. Extracting temporal information from electronic patient records. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium Proceedings*. [S.l.], 2012. v. 2012, p. 542. Cited on page 18.
- LIU, Z. et al. Cmedtex: A rule-based temporal expression extraction and normalization system for chinese clinical notes. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium Proceedings*. [S.l.], 2016. v. 2016, p. 818. Cited on page 16.
- LIU, Z. et al. Temporal indexing of medical entity in chinese clinical notes. *BMC medical informatics and decision making*, BioMed Central, v. 19, n. 1, p. 17, 2019. Cited on page 47.
- MACAVANEY, S.; COHAN, A.; GOHARIAN, N. Guir at semeval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. [S.l.: s.n.], 2017. p. 1024–1029. Cited on page 49.
- MADKOUR, M.; BENHADDOU, D.; TAO, C. Temporal data representation, normalization, extraction, and reasoning: A review from clinical domain. *Computer methods and programs in biomedicine*, Elsevier, v. 128, p. 52–68, 2016. Cited on page 25.
- MANFREDI, G. et al. Heidelttime at eventi: Tuning italian resources and addressing timeml’s empty tags. In: *Proceedings of the Forth International Workshop EVALITA*. [S.l.: s.n.], 2014. p. 39–43. Cited on page 45.
- MANI, I. et al. Machine learning of temporal relations. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. [S.l.], 2006. p. 753–760. Cited on page 20.
- MANNING, C. D.; SCHÜTZE, H. et al. *Foundations of statistical natural language processing*. [S.l.]: MIT Press, 1999. Quoted 2 times on pages 31 and 32.
- MENG, Y.; RUMSHISKY, A.; ROMANOV, A. Temporal information extraction for question answering using syntactic dependencies in an lstm-based architecture. *arXiv preprint arXiv:1703.05851*, 2017. Cited on page 14.
- MOHARASAN, G.; HO, T.-B. Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research*, Springer, p. 1–25, 2019. Cited on page 47.

- MOHARASAR, G.; HO, T. B. A semi-supervised approach for temporal information extraction from clinical text. In: IEEE. *Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2016 IEEE RIVF International Conference on*. [S.l.], 2016. p. 7–12. Quoted 3 times on pages 16, 47, and 118.
- MORICEAU, V.; TANNIER, X. French resources for extraction and normalization of temporal expressions with heideltme. *extraction*, v. 26, 2013. Cited on page 45.
- MOTA, C.; SANTOS, D. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. [S.l.]: Linguatca, 2008. Cited on page 18.
- NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. *Linguisticae Investigationes*, John Benjamins publishing company, v. 30, n. 1, p. 3–26, 2007. Cited on page 22.
- NZALI, M. D. T.; TANNIER, X.; NÉVÉOL, A. Automatic extraction of time expressions accross domains in french narratives. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2015. p. 492–498. Cited on page 16.
- PIVOVAROV, R.; ELHADAD, N. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, Oxford University Press, v. 22, n. 5, p. 938–947, 2015. Cited on page 15.
- PUSTEJOVSKY, J. et al. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, v. 3, p. 28–34, 2003. Quoted 6 times on pages 14, 25, 27, 28, 59, and 72.
- PUSTEJOVSKY, J. et al. Temporal and event information in natural language text. *Language resources and evaluation*, Springer, v. 39, n. 2, p. 123–164, 2005. Quoted 2 times on pages 14 and 25.
- PUSTEJOVSKY, J.; VERHAGEN, M. Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. [S.l.], 2009. p. 112–116. Cited on page 20.
- RAZALI, N. M.; WAH, Y. B. et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, v. 2, n. 1, p. 21–33, 2011. Cited on page 85.
- REEVES, R. M. et al. Detecting temporal expressions in medical narratives. *International journal of medical informatics*, Elsevier, v. 82, n. 2, p. 118–127, 2013. Cited on page 20.
- RIM, K. Mae2: Portable annotation tool for general natural language use. In: *Proceedings of 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. [S.l.: s.n.], 2016. p. 75–80. Cited on page 62.
- ROBERTS, K.; RINK, B.; HARABAGIU, S. M. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, BMJ Publishing Group BMA House, Tavistock Square, London, WC1H 9JR, v. 20, n. 5, p. 867–875, 2013. Cited on page 14.

- SARATH, P.; MANIKANDAN, R.; NIWA, Y. Hitachi at semeval-2017 task 12: System for temporal information extraction from clinical notes. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. [S.l.: s.n.], 2017. p. 1005–1009. Cited on page 14.
- SAURI, R. et al. *TimeML Annotation Guidelines Version 1.2. 1*. [S.l.]: January, 2006. Quoted 5 times on pages 28, 59, 72, 122, and 134.
- SCHILDER, F. Extracting meaning from temporal nouns and temporal prepositions. *ACM Transactions on Asian Language Information Processing (TALIP)*, ACM, v. 3, n. 1, p. 33–50, 2004. Cited on page 14.
- SKUKAN, L.; GLAVAŠ, G.; ŠNAJDER, J. Heideltime. hr: extracting and normalizing temporal expressions in croatian. In: *Proceedings of the 9th Slovenian Language Technologies Conferences (IS-LT 2014)*. [S.l.: s.n.], 2014. p. 99–103. Cited on page 45.
- STRÖTGEN, J. *Domain-sensitive temporal tagging for event-centric information retrieval*. Tese (Doutorado), 2015. Quoted 9 times on pages 26, 52, 92, 122, 123, 124, 134, 135, and 136.
- STRÖTGEN, J.; GERTZ, M. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 5th International Workshop on Semantic Evaluation*. [S.l.], 2010. p. 321–324. Quoted 3 times on pages 19, 45, and 52.
- STRÖTGEN, J.; GERTZ, M. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, Springer, v. 47, n. 2, p. 269–298, 2013. Quoted 6 times on pages 7, 34, 35, 36, 41, and 52.
- STRÖTGEN, J.; GERTZ, M. A baseline temporal tagger for all languages. In: *EMNLP*. [S.l.: s.n.], 2015. v. 15, p. 541–547. Quoted 3 times on pages 19, 45, and 52.
- STRÖTGEN, J.; GERTZ, M. Domain-sensitive temporal tagging. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, v. 9, n. 3, p. 1–151, 2016. Quoted 13 times on pages 7, 9, 14, 22, 23, 24, 25, 27, 31, 32, 33, 45, and 52.
- STUBBS, A. Mae and mai: lightweight annotation and adjudication tools. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 5th Linguistic Annotation Workshop*. [S.l.], 2011. p. 129–133. Cited on page 62.
- STYLER, I.; WILLIAM, F. et al. *THYME annotation guidelines*. [S.l.], 2014. Quoted 10 times on pages 44, 72, 122, 123, 124, 126, 134, 135, 136, and 138.
- SUN, W.; RUMSHISKY, A.; UZUNER, O. Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, Elsevier, v. 46, p. S5–S12, 2013. Cited on page 44.
- TANG, B. et al. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, BMJ Publishing Group BMA House, Tavistock Square, London, WC1H 9JR, v. 20, n. 5, p. 828–835, 2013. Cited on page 47.
- TISSOT, H. et al. Normalisation of imprecise temporal expressions extracted from text. *Knowledge and Information Systems*, 2019. Cited on page 18.

- TISSOT, H. et al. Ufprsheffield: Contrasting rule-based and support vector machine approaches to time expression identification in clinical tempeval. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. [S.l.: s.n.], 2015. p. 835–839. Cited on page 48.
- TISSOT, H.; PESCHL, G.; FABRO, M. D. D. Fast phonetic similarity search over large repositories. In: SPRINGER. *International Conference on Database and Expert Systems Applications*. [S.l.], 2014. p. 74–81. Cited on page 19.
- TISSOT, H. et al. Analysis of temporal expressions annotated in clinical notes. In: *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*. [S.l.: s.n.], 2015. Cited on page 72.
- TISSOT, H. C. Normalisation of imprecise temporal expressions extracted from text. 2016. Quoted 5 times on pages 19, 42, 55, 125, and 137.
- TOURILLE, J. et al. Limsi-cot at semeval-2017 task 12: Neural architecture for temporal information extraction from clinical narratives. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *11th International Workshop on Semantic Evaluation (SemEval-2017)*. [S.l.], 2017. p. 597–602. Cited on page 49.
- UZZAMAN, N.; ALLEN, J. F. Event and temporal expression extraction from raw text: First step towards a temporally aware system. *International Journal of Semantic Computing*, World Scientific, v. 4, n. 04, p. 487–508, 2010. Quoted 2 times on pages 14 and 47.
- VELUPILLAI, S. et al. Blulab: Temporal information extraction for the 2015 clinical tempeval challenge. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. [S.l.: s.n.], 2015. p. 815–819. Cited on page 48.
- VERHAGEN, M. et al. Semeval-2007 task 15: Tempeval temporal relation identification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 4th International Workshop on Semantic Evaluations*. [S.l.], 2007. p. 75–80. Quoted 2 times on pages 43 and 44.
- VERHAGEN, M. et al. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, Springer, v. 43, n. 2, p. 161–179, 2009. Quoted 3 times on pages 14, 20, and 22.
- VERHAGEN, M. et al. Automating temporal annotation with tarsqi. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*. [S.l.], 2005. p. 81–84. Cited on page 45.
- VERHAGEN, M. et al. Semeval-2010 task 13: Tempeval-2. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 5th international workshop on semantic evaluation*. [S.l.], 2010. p. 57–62. Cited on page 43.
- VIANI, N. et al. Time expressions in mental health records for symptom onset extraction. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. [S.l.: s.n.], 2018. p. 183–192. Quoted 4 times on pages 44, 92, 112, and 113.
- ZHOU, X.; WANG, L. Research on temporal expression recognition: A case study of uyghur. Cited on page 16.