Temporal Tagging of Noisy Clinical Texts in Brazilian Portuguese

Rafael Faria de Azevedo[1]
João Pedro Santos Rodrigues
Mayara Regina da Silva Reis
Claudia Maria Cabral Moro[2]
Emerson Cabrera Paraiso[1]

[2]Programa de Pós-Graduação em Tecnologia em Saúde (PPGTS)
[1]Programa de Pós-Graduação em Informática (PPGIa)

Summary

Introduction

Motivation

- Make temporal tagging of noisy real clinical texts in Brazilian Portuguese
- Creation of the patient timeline
- Summarizing reports of a long time chronic patient

Goal

Create a method to make the temporal tagging task of noisy clinical texts written in Brazilian Portuguese

Contribution

Method to extract and normalize well-written and noisy temporal expressions from clinical real data in Brazilian Portuguese

Backgroud

Temporal Expressions

- E.g.: 2010, yesterday, discharge day etc.

News Domain (Commonly studied)

- Correctly written
- English
- Competition datasets

Temporal expressions are part of other processes like:

- Question answering
- Search

Example

Fulano , 52 anos # sida em uso de abacavir+efavirenz+lamivudina (último cd4 maio/07 875) - em acompanhamento no ambulatório de infecto hcpa # hcv+ (...) # fístula pleuro-cutânea no local de bx pleural - baar na secreção + (não realizada cultura para micobactéria) - rhz por 9 meses

Examples

- 15:40 1º DI Paciente vitima de... As 10: 00
- 15:40 1º dia de internação Paciente vitima de... As 10:00

- procedimento cirúrgico em 07/01 as07h. Enf xxxxxxxx
- procedimento cirúrgico em 07/01 às 07h. Enf xxxxxxxx

- BEXIGA CHEIA E SEMPRE NO FINM DO DIA...
- BEXIGA CHEIA E SEMPRE NO FIM DO DIA...

Basic concepts

Temporal Tagging = Extraction + Normalization

Extraction

<TIMEX3 type="DURATION" value="P38W1D">**38 sem e 1 dia**</TIMEX3>

Normalization

<TIMEX3 **type="DURATION" value="P38W1D"**>38 sem e 1 dia</TIMEX3>

Basic concepts

Temporal Expression types: DATE, TIME, DURATION, SET

DATE

<TIMEX3 type="**DATE**" value="**2006-12**">**dezembro . 06**</TIMEX3>

TIME

<TIMEX3 tid="t224" type="**TIME**" value="**12:30**">**12:30**</TIMEX3>

Basic concepts

Temporal Expression types: DATE, TIME, DURATION, SET

DURATION

<TIMEX3 tid="t48“ type="**DURATION**" value="**P52Y**">**52 anos**</TIMEX3>

SET

<TIMEX3 type="**SET**" value="**P1D**" freq="**3X**">**3 x ao dia**</TIMEX3>

Basic concepts

HeidelTime[a]

[a] https://github.com/HeidelTime/heideltime

(Part **1**) **#38 sems 1 dia (s)**, 38 **SEM** e 1 dia (s), **38sem e1dia(s)**

(Part **2**) **RULENAME**="clinical_duration_r17b",
EXTRACTION="([\d]+)(\s)?(%**reUnitAbbrev**)(\s)?e?(\s)?([\d]+)(\s)?(%reUnit)",
NORM_VALUE="Pgroup(1)%**normUnit4DurationAbbrev**(group(3))group(7)%normUnit4Duration(group(9))"

(Part **3**) # <TIMEX3 tid="t12" type="DURATION" value="P38W1D">**38 sems 1 dia**</TIMEX3> (s) ,
<TIMEX3 tid="t13" type="DURATION" value="P38W1D">**38 sem e 1 dia**</TIMEX3> (s) ,
<TIMEX3 tid="t14" type="DURATION" value="P38W1D">**38sem e1dia**</TIMEX3> (s)

Related Works

Related Works

- Approaches (Kreimeyer, K., et al. 2017):
  - 1st - Rule-based approach
  - 2nd - Hybrid approach
  - 3rd - Machine learning approach
- News domain (Portuguese)
  - (Costa, F., Branco, A. 2012)

Proposed Method

Dataset

Fulano , 52 anos # sida em uso de abacavir+efavirenz+lamivudina (último cd4 maio/07 875) - em acompanhamento no ambulatório de infecto hcpa # hcv+ (...) # fístula pleuro-cutânea no local de bx pleural - baar na secreção + (não realizada cultura para micobactéria) - rhz por 9 meses

Dataset

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE TimeML SYSTEM "TimeML.dtd">
<TimeML>
    Fulano ,
    <TIMEX3 value="P52Y" type="DURATION" tid="t41">52 anos</TIMEX3>
    # sida em uso de abacavir+efavirenz+lamivudina ( último cd4
    <TIMEX3 value="2007-05" type="DATE" tid="t8">maio/07</TIMEX3>
    875 ) - em acompanhamento no ambulatório de infecto hcpa # hcv+
    (omitted by the author) # fístula pleuro-cutânea no local de bx
    pleural - baar na secreção + ( não
    <TIMEX3 value="PAST_REF" type="TIME" tid="t22">realizada</TIMEX3>
    cultura para micobactéria ) - rhz por
    <TIMEX3 value="P9M" type="DURATION" tid="t44">9 meses</TIMEX3>
</TimeML>
```

Dataset

- Data written in Portuguese from Brazilian hospitals
- Collected between 2002 and 2007
- 1000 unlabeled clinical texts
  - Unlabeled
    - 870 texts
  - Labeled
    - 130 texts

Dataset

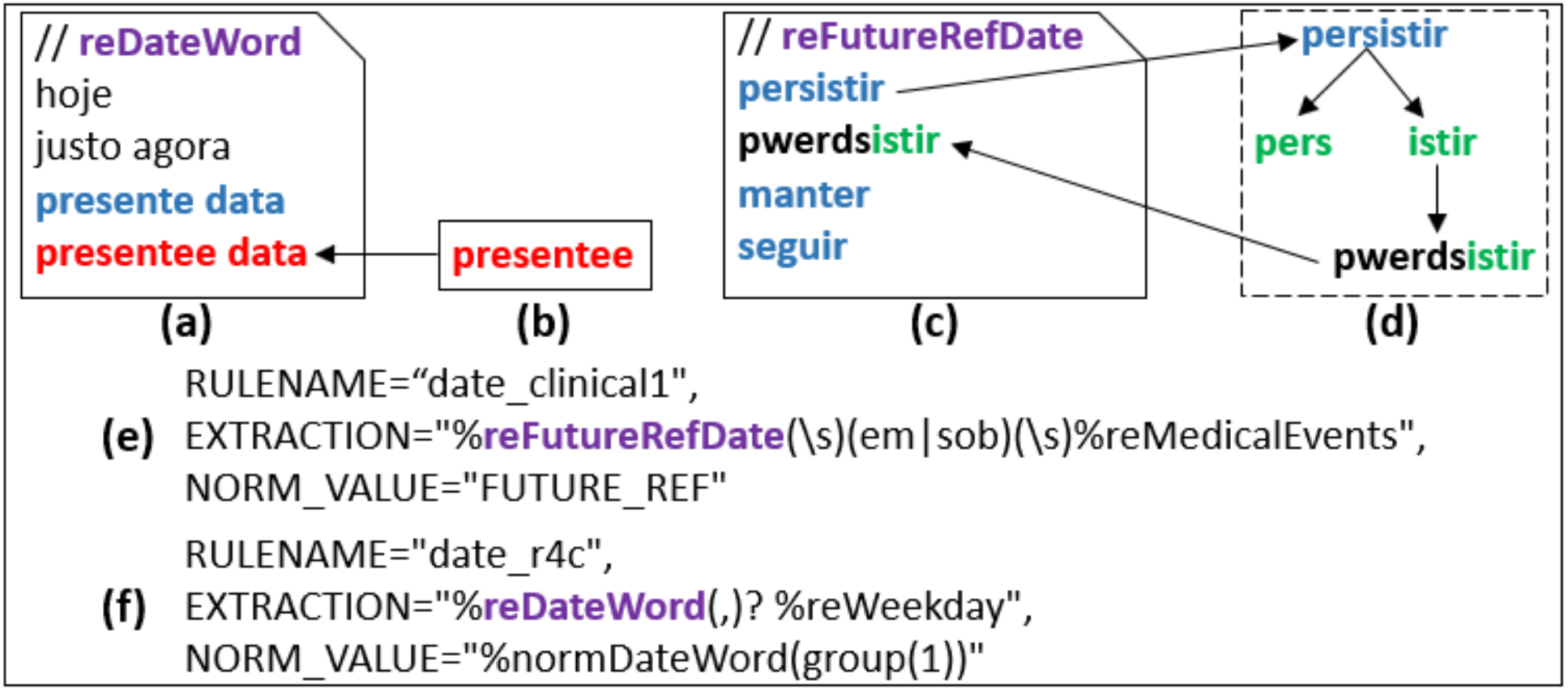
- Labeled
  - Annotated by:
    - Nurse assistant
    - Student in computational linguistics
    - Kappa was 75.2 (normalization)
  - 100 texts to train (77%)
  - 30 texts to test (23%)

Preprocessing step

- Lowercase
- Corrected some errors

Noisy Temporal Exps	Preprocessed
yesterday morning	yesterday morning

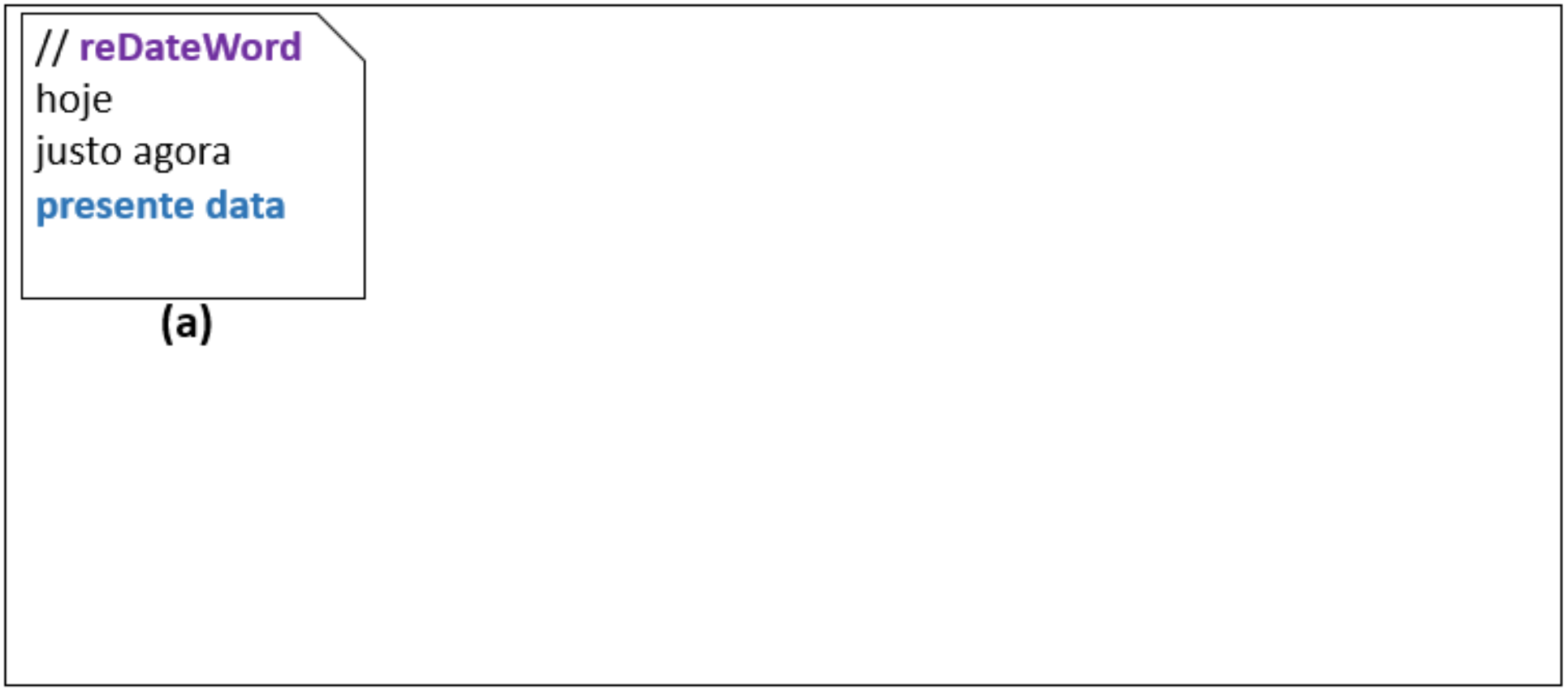
Processing step - Method overview



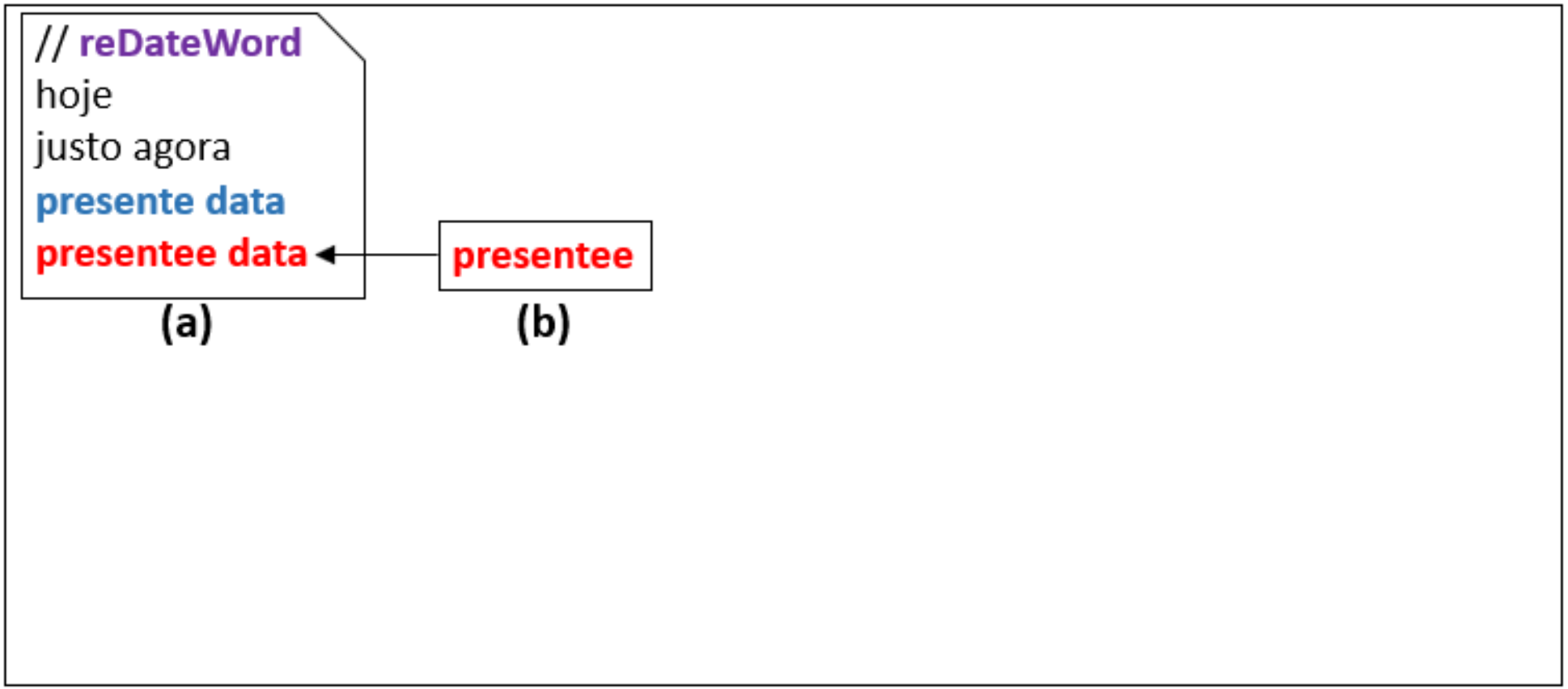
(e)
RULENAME="date_clinical1",
EXTRACTION="%reFutureRefDate(\s)(em|sob)(\s)%reMedicalEvents",
NORM_VALUE="FUTURE_REF"

(f)
RULENAME="date_r4c",
EXTRACTION="%reDateWord(,)? %reWeekday",
NORM_VALUE="%normDateWord(group(1))"

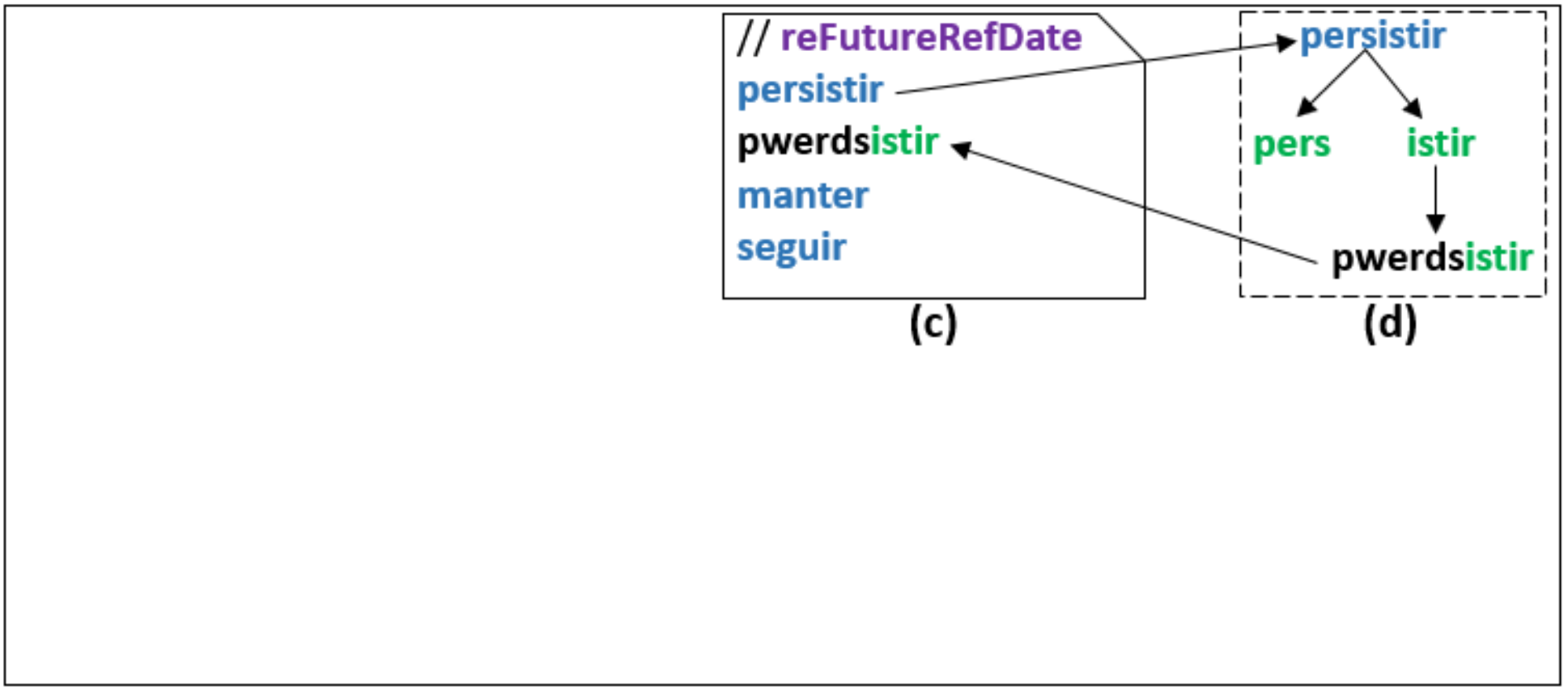
Processing step - 1st approach - (Correct)



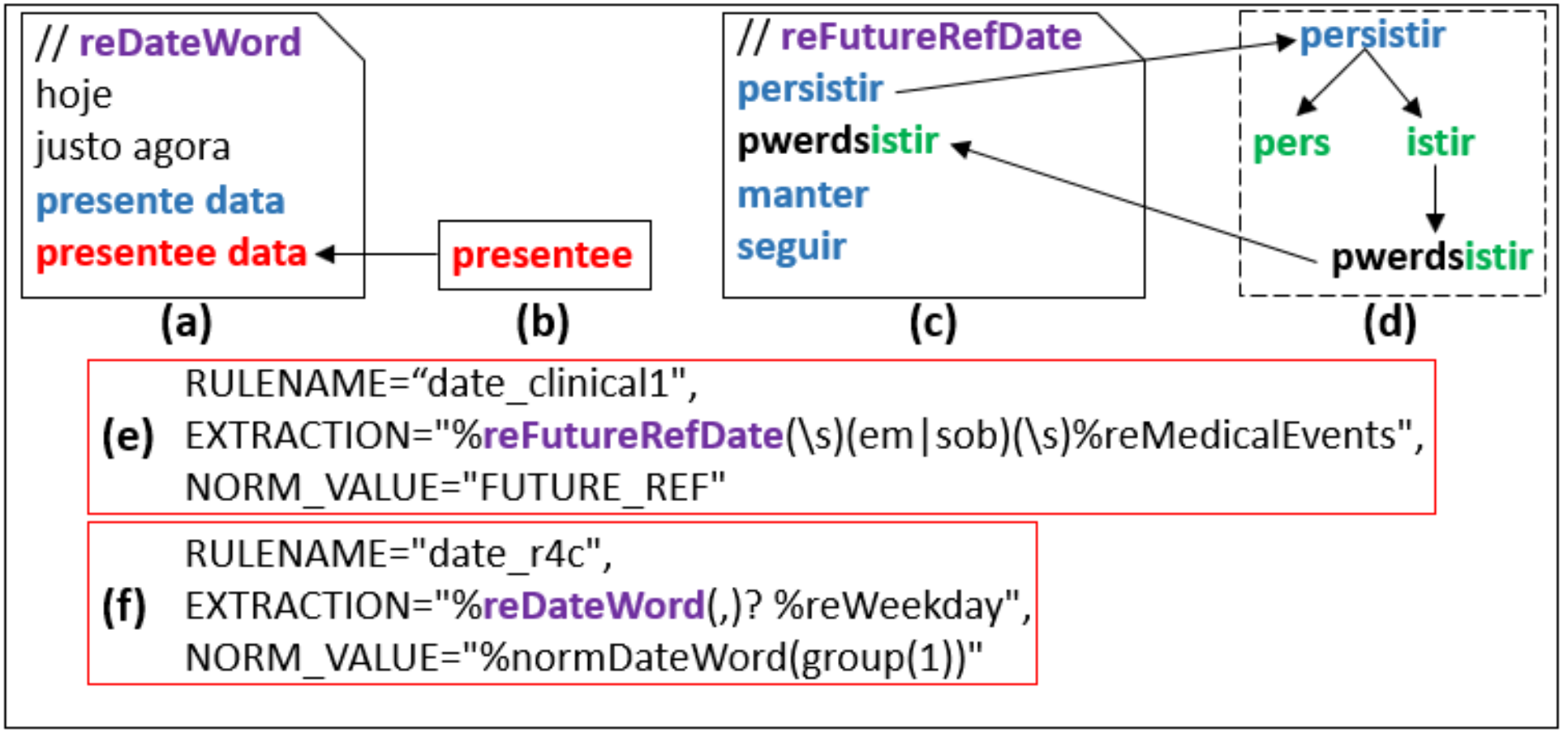
Processing step - 2nd approach - (Noisy)



Processing step - 3rd approach - (N-gram)



Processing step - Whole method



Experiments and Results

Experiments

Step	Metric	Experiments			
		Baseline (1)	Correct (2)	Noisy (3)	N-Gram (4)
Extraction	precision	79.78	**95.58**	93.18	94.52
	recall	17.27	68.37	76.40	**83.94**
	F1 score	28.40	79.72	83.96	**88.92**
Normalization	precision	77.53	91.50	92.58	**93.42**
	recall	16.79	65.45	75.91	**82.97**
	F1 score	27.60	76.31	83.42	**87.89**

Results

- Statistical tests applied:
  - Shapiro-Wilk
  - Friedman
  - Dunn-Bonferroni
  - Wilcoxon
  - Significance of 0.05
- Statistical difference:
  - Experiment 2 and experiment 4 were statistically different

Results

- N-gram strategy found:
  - 28 noisy temporal expressions

Noisy Temporal Exps	Correct form Portuguese	Correct form English
aanterior	anterior	previous
seginte	seguinte	following/next
acompanham,ento	acompanhamento	accompany/tracking

Results

- Temporal expressions missed by HeidelTime

Noisy Temporal Exps	Correct form Portuguese	Correct form English
2m	2 meses	2 months
1a5m	1 ano e 5 meses	1 year and 5 months
pó-op	pós-operatório	postoperative

Results

- A problematic temporal expression
  - 02/07
- Annotators found that 7.69% of temporal expressions were noisy

Conclusion

Conclusion

- Results are similar to other works made in other languages, especially within the clinical domain
- None of the other works coped with noisy temporal expressions in the extraction and normalization steps using a rule-based approach for Portuguese

Future work

- A new dataset is being finished and will be made available to the community
- Our proposed method will be tested in the new dataset
- We intend to used the proposed method along with a hybrid approach

Acknowledgment

We would like to thank the Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES) and PUCPR for their financial support.

Thanks for your attention!

rafael@ppgia.pucpr.br
jpsanr@gmail.com
mayara.reis@outlook.com
c.moro@pucpr.br
paraiso@ppgia.pucpr.br

Temporal Tagging of Noisy Clinical Texts in Brazilian Portuguese

Rafael Faria de Azevedo[1]
João Pedro Santos Rodrigues
Mayara Regina da Silva Reis
Claudia Maria Cabral Moro[2]
Emerson Cabrera Paraiso[1]

[2]Programa de Pós-Graduação em Tecnologia em Saúde (PPGTS)
[1]Programa de Pós-Graduação em Informática (PPGIa)