A Novel Method for Identifying Continuity of Care in Hospital Discharge Summaries

Lucas Emanuel Silva e Oliveira[1], Andréia Cristina de Souza[1],
Percy Nohama[1,2], and Claudia Maria Cabral Moro[1]

[1] Polytechnic School, Pontifical Catholic University of Paraná, Curitiba, Paraná, Brazil
[2] Post-Graduate Program of Electrical and Industrial Engineering, Federal Technological University of Paraná, Curitiba, Paraná, Brazil

Abstract— **Discharge summary is one of the most important clinical narratives because it provides essential information about the patient, including the continuity of care. Due to its free-form writing and the lack of consensus on its essential content, the identification of data contained on the text becomes a difficult task. In this project, we propose a rule-based method to verify the presence of information related to the continuity of care in Brazilian Portuguese texts, applying Natural Language Processing (NLP) techniques with an open-source tool named CoGrOO, and based on an annotated medical corpus. After the experiments, four rules were defined and applied on 200 summaries to identify the existence of the content "continuity of care". This process had resulted in a Precision value of 83%, Recall value of 69% and F-Measure value of 76% related to algorithm evaluation.**

Keywords— **Natural Language Processing, Discharge Summaries, Clinical Narratives.**

I. INTRODUCTION

The information produced during care provided to patients are recorded in the medical records, some of them in the form of clinical narratives and others in structured fields. One of these narratives is the discharge summary, a brief report of the patient's hospitalization, which presents important information for his continuity of care. [1]

The systematic identification of information in discharge summaries is currently a great challenge, because its preparation is not performed in the form of structured fields, but in free-form writing, using natural language. Moreover, there is still no consensus on the adequate and essential content that must be present in discharge summaries, so each professional prepares the summary differently, and often fails in recording important data, which may result in adverse events to the patients and difficulties to the medical staff that will continue the treatment [2].

There are several studies that focus on the recovery and identification of information in clinical narratives. Bui et al. [3] developed a tool for automated extraction of data related to resistance of HIV virus in the application of drugs, using Natural Language Processing (NLP). Xu et al. [4] used NLP to extract information related to drugs in clinical narratives. Bulegon [5] used an open-source tool called CoGrOO to extract diagnoses reported in discharge summaries, and subsequently perform its mapping to ICD-10.

In this paper, we present a research whose goal was to develop a rule-based method to identify information regarding continuity of patient care. Therefore, it shows the summaries that does not contain this kind of information. Also, it was intended to develop an automated tool to run the algorithm named IRDischarge.

II. MATERIALS AND METHODS

The method was based on the Bulegon's research [5], which identifies diagnoses present in discharge summaries using morphological rules. The rules were created specifically for each disease, and each rule contains three terms, the relevant information is always the central term, in that case, the disease itself.

The database used for the development of the project was provided by the Hospital of Porto Alegre (HCPA). A total of 5617 discharge summaries from patients of cardiology service were discharged from June 2002 to May 2007. Among all summaries, we chose only those who had some data concerning continuity of patient care, and among these, they were randomly selected 110 of them.

For the processing of texts, some tools were analyzed [6,7,8], and was chosen CoGrOO[8], an open-source tool that implements the most common techniques of NLP, including: tokenizer, sentence detector, name finder, POS-tagging, among others.

The tagging (part-of-speech tagging or POS-tagging) is one of the techniques used in NLP for obtaining morphological value of words, and from this information Bulegon's rules were defined [5]. The acronyms generated in tagging are part of a tagset (tags dictionary) called VISL Portuguese [9], which is the default tagset of CoGrOO 3.0.5 version.

Then, it was necessary the tagging of texts for morphological analysis and the identification of patterns that would indicate the presence of continuity of care in their content, as presented on Table 1. The meanings of the tags are shown in Table 2.

The POS-Tagging is based on statistical models for classifying the words found in the text, and by default, these

I. Lacković et al. (eds.), *The International Conference on Health Informatics*, IFMBE Proceedings 42,
DOI: 10.1007/978-3-319-03005-0_72,

statistical models are trained in a journalistic corpus: we use texts with common vocabulary in different contexts. Since this research involves clinical narratives, it was necessary to use a consistent corpus with the application domain, so that the accuracy rate was higher in tagging process. The corpus used in the sequence was the same annotated medical corpus in the work of Peters et al. [10].

Table 1: Examples of Portuguese sentences after the POS-Tagging.

Sentence	Sentence after the POS-Tagging
Paciente com sarcoma sinovial, interna para exames de reavaliação com TC	Paciente_N_M_S com_PRP sarcoma_N_F_S sinovial_ADJ_F_S interna_V_PR_3S_IND_VFIN para_PRP exames_N_M_P de_PRP reavaliação_N_F_S
É encaminhado para ambulatório do médico assistente	É_V_PR_3S_IND_VFIN encaminhado_V_PCP_M_S para_PRP ambulatório_N_M_S de_PRP o_DET_M_S médico_N_M_S assistente_ADJ_M_S

Table 2: Meaning of the mophological labels.

Tag	Meaning
ADJ_F_S	Adjective Feminine Single
DET_M_S	Article Male Single
N_F_S	Noum Feminine Single
N_M_P	Noum Male Plural
N_M_S	Noum Male Single
PRP	Preposition
V_PR_3S_IND_VFIN	Verb Present Indicative Tense Third Person Single

With the summaries already tagged, it was performed a manual analysis to evaluate the accuracy rate of the POS-Tagger. It has been found that even making use of the annotated medical corpus, some words´ classification had been done incorrectly. So, some processes were performed in a pre-processing stage, before sending the texts to the POS-Tagger.

As acronyms are common in clinical narratives and they hamper NLP execution, the acronym expansion process was performed, replacing them by their full meaning, using a list created based on the same corpus used in this research.

After this expansion process, the whole text was transformed in lowercase to avoid reducing the success-rate of POS-tagging because the corresponding word could be absent in the trained model. It becomes necessary inasmuch narratives did not follow the rules of formal writing generally, where the sentences begin with uppercase letters and the others words in lowercase.

The words with high occurrence in the text and without significant semantic value, known as stop-words, were also removed. Only the articles were considered stop-words in that context. In addition, it was also removed special characters without meaning, such as "#", "@", and "*".

Once the problems were corrected in the stage of tagging, summaries analysis was done looking for parts in which the information of healthcare continuity was present. Then, patterns of tags that validate the presence of those data in the discharge summary were chosen.

Unlike Bulegon work [5], the elaboration or rules were not limited to three terms, thus enabling the creation of broader and more specific patterns. The continuity of care related information is characterized by parts of sentences, not only by some specific word, like the central term of Bulegon method. Furthermore the method is intended only to verify the presence o+f information, rather than extract it, discarding the need for a central term.

The definition of the rules was done by manual analysis and iteratively, checking at the end of each iteration their effectiveness and coverage in the texts. At the end of each iteration, we selected 200 new summaries randomly to re-evaluate the remaining rules, resulting, at the end of the process, on four rules.

During the process of defining rules, it was realized that the information regarding the continuity of care were generally arranged in the last sentence of the summary, which led us to reverse the order verification of sentences, looking from the last to the first sentence .

To evaluate the effectiveness of the developed method and the defined rules, we had selected 200 discharge summaries randomly, and run the algorithm for detecting the information.

For assessment of the four generated rules, parameters Precision, Recall and F-Measure were computed according to Equations 1, 2 and 3).

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives}+\text{False positives}} \quad (1)$$

$$\text{Recall} = \frac{\text{True negatives}}{\text{True negatives}+\text{False negatives}} \quad (2)$$

$$\text{F} - \text{measure} = \frac{\text{precision}*\text{recall}}{\text{precision}+\text{recall}} \quad (3)$$

III. RESULTS

The proposed approach requires the implementation of the process according the following sequence:

- Pre-processing on IRDischarge
  - Acronym expansion
  - Special characters removal
  - Lowercase Normalization
  - Stop-words removal
- Processing on Cogroo
  - Sentence detector
  - Tokenizer
  - Name Finder
  - Preposition expansion
  - POS-Tagging
- Processing on IRDischarge
  - Sentence inversion
  - Rules indentification on text.

For assessment of the four generated rules (Table 3), the parameters Precision, Recall and F-Measure were computed and their results are presented on Table 4.

Table 3: Generated rules.

Rules
[V_INF][N_M_S][PRP]
[V_INF][N_M_S][ADJ_M_S]
[V_INF][PRP][N_M_S]
[V_INF][N_M_S]

Table 4: Evaluation methods and values.

Measure method	Value
Precision	83%
Recall	69%
F-Measure	76%

IV. DISCUSSION

Although Bulegon [5] developed a method that also uses discharge summaries, it required several changes in the proposed methodology to suit the verification of content related to continuity of care. Among them, the reading of sentences in reverse mode because, unlike the diagnostic information that is usually located at the beginning of the text, continuity of care is positioned at the end.

The composition of patterns of rules was another feature that was altered because the methods differ in the handling of the information; while Bulegon [5] needs to extract information from diagnostics, the novel approach requires only the inspection of the presence or absence of data related to continuity of care.

Pompeo et al. [11] indicate the difficulty in retrieving information if they are not registered correctly. This problem has provoked the major cases of false positives in the analysis of summaries. The most correct mode to write and store information is the structured format. It contains text prepared properly with well-defined and consistent data. [12]

Well-developed approaches for identifying information in discharge summaries, and integrated in tools built into the day-to-day of the healthcare professionals would be of great interest to all professionals who prepare those documents, reducing data loss and problems on continuity of care of the patient. [13].

V. CONCLUSION

The proposed method, using only four rules, may identify one of the most important information described in discharge summaries, the continuity of care. If the approach is implemented as a part of a tool inside the hospital environment, it will can support the correct writing of discharge summaries and improve the continuity of patient care.

For future studies, the method can also be adapted to identify other essential information in discharge summary, as diagnostic and therapeutic procedures performed and medications recommended for discharge. In addition we can create a semi-assisted algorithm for rules creation.

ACKNOWLEDGMENTS

The authors are grateful to CNPq for funding this research and to Mariza Machado Kluck for the database containing the discharge summaries of HCPA.

REFERENCES

1. Grossman E, Cardoso MHCA. As narrativas em medicina: contribuições à prática clínica e ao ensino médico. Revista Brasileira de Educação Médica. v.30 n.1, 2006.
2. Kripalani S, Lefevre F, Phillips CO, Williams MV, Basaviah P, BAKER DW. Deficits in communication and in-formation transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA. vol. 297, n.8. fev. 2007.
3. Bui QC, Nualláin BO, Boucher CA, Sloot PM. Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics. Fevereiro 23;11:101, 2010.
4. Xu H, Stetson PD, Friedman C. A Study of Abbreviations in Clinical Notes. AMIA Annual Symposium Proceedings, 2007.
5. Bulegon, Hugo. Identificação de diagnósticos contidos em narrativas clínicas e mapeamento para a classificação internacional de doenças. 2011. 101f. Dissertação (Mestrado em Tecnologia em Saúde) – PUCPR, Curitiba, 2011.
6. OpenNLP, retrieved from http://opennlp.sourceforge.net, last access on 10/12/2012.
7. Natural Language Toolkit, retrieved from http://nltk.org, last access on 10/12/2012.

8. CoGrOO, retrieved from http://cogroo.sourceforge.net, last access on 10/12/2012.
9. VISL Portuguese, retrieved from http://beta.visl.sdu.dk, last access on 10/12/2012.
10. Peters AC, Oleynik M, Pacheco EJ, Barra C MCM, Schulz SP, Nohama P. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. In: Anais do XII Congresso Brasileiro de Informática em Saúde, 2010. p. p1-p6.
11. Pompeo DA, Pinto MH, Cesarino CB, Araújo RRDF, Poletti NAA. Atuação do enfermeiro na alta hospitalar: reflexões a partir dos relatos de pacientes. Acta Paulista de Enfermagem. vol. 20 no.3. São José do Rio Preto, São Paulo, 2007.
12. Long W. Lessons extracting diseases from discharge summaries. AMIA, Symposium Proceedings. Massachusetts Institute of Technology, Cambridge, USA. p.478-482. Published online. 2007.
13. Petruci FR; Benefícios da contra referência na alta hospitalar para equipe da atenção básica. Monografia apresentada à INDEP – Instituto de Ensino e Capacitação e Pós Graduação, como parte dos requisitos para obtenção do título de Especialista em Saúde Pública com ênfase em Estratégia de Saúde da Família. Assis-SP, 2010.

Address for correspondence

Author: Percy Nohama
Institute: Pontifícia Universidade Católica do Paraná
Street: Rua Imaculada Conceição 1155, CEP 80215-901
City: Curitiba
Country: Brazil
Email: percy.nohama@gmail.com