

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE

LUCAS EMANUEL SILVA E OLIVEIRA

**IDENTIFICAÇÃO DE CONTINUIDADE DE CUIDADO EM SUMÁRIOS
DE ALTA HOSPITALAR**

CURITIBA

2014

LUCAS EMANUEL SILVA E OLIVEIRA

**IDENTIFICAÇÃO DE CONTINUIDADE DE CUIDADO EM SUMÁRIOS
DE ALTA HOSPITALAR**

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde, da Pontifícia Universidade Católica do Paraná, como requisito parcial para obtenção do título de Mestre em Tecnologia em Saúde.

Área de concentração: Tecnologia em Saúde

Orientador: Prof^a. Dr^a. Claudia Maria Cabral Moro Barra.

Co-orientador: Prof^o. Dr. Percy Nohama

**CURITIBA
2014**

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

Oliveira, Lucas Emanuel Silva e
O48i Identificação de continuidade de cuidado em sumários de alta hospitalar /
2014 Lucas Emanuel Silva e Oliveira ; orientadora, Claudia Maria Cabral Moro
Barra ; co-orientador, Percy Nohama. – 2014.
69 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná,
Curitiba, 2014

Bibliografia: f. 61-69

1. Alta do paciente. 2. Registros médicos - Processamento de dados.
3. Pacientes - Cuidados médicos. 4. Processamento de linguagem natural
(Computação). I. Barra, Cláudia Maria Cabral Moro, 1969-. II. Nohama, Percy.
III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação
em Tecnologia em Saúde. IV. Título.

CDD 20. ed. – 610.28



Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Tecnologia em Saúde

**ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

DEFESA DE DISSERTAÇÃO Nº 184

ÁREA DE CONCENTRAÇÃO: TECNOLOGIA EM SAÚDE

Aos 25 dias do mês de março de 2014, no auditório John Henry Newman, realizou-se a sessão pública de Defesa da Dissertação: "Identificação de continuidade de cuidado em sumários de alta hospitalar", apresentada pelo aluno Lucas Emanuel Silva e Oliveira, sob orientação do Prof^a. Dr^a. Claudia Maria Cabral Moro Barra e coorientação do Prof. Dr. Percy Nohama, como requisito parcial para a obtenção do título de Mestre em Tecnologia em Saúde, perante uma Banca Examinadora composta pelos seguintes membros:

Prof ^a . Dr ^a . Claudia Maria Cabral Moro Barra, PUCPR (Orientador e presidente)	<u>[assinatura]</u> (assinatura)	<u>APROVADO</u> (Aprov/Reprov.)
Prof. Dr. Emerson Paraíso, PUCPR (Examinador)	<u>[assinatura]</u> (assinatura)	<u>Aprov.</u> (Aprov/Reprov.)
Prof. Dr. Marco Antônio Gutierrez, Incor HCFMUSP (Examinador)	<u>[assinatura]</u> (assinatura)	<u>APROVADO</u> (Aprov/Reprov.)
Prof ^a . Dr ^a . Eneida Abrantes Mendonça, University of Wisconsin (Examinador)	<u>[assinatura]</u> (assinatura)	<u>APROVADO</u> (Aprov/Reprov.)

Início: 15:15 Término: 17:35

Conforme as normas regimentais do PPGTS e da PUCPR, o trabalho apresentado foi considerado APROVADO (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora.

Observações: Revisão do texto, estrutura das frases,
referências e citações referente ao conteúdo
Atendendo sugestões indicadas pela banca

O aluno está ciente que a homologação deste resultado está condicionada: (I) ao cumprimento integral das solicitações da Banca Examinadora, que determina um prazo de 90 dias para o cumprimento dos requisitos; (II) entrega da dissertação em conformidade com as normas especificadas no Regulamento do PPGTS/PUCPR; (III) entrega da documentação necessária para elaboração do Diploma.

ALUNO: Lucas Emanuel Silva e Oliveira [assinatura]
(assinatura)

[assinatura]
Prof^a. Dr^a. Marcia Regina Cúbas,
Coordenadora do PPGTS PUCPR

AGRADECIMENTOS

Primeiramente agradeço a Prof^a. Dr^a. **Cláudia Moro**, pela sua maneira peculiar e atenciosa de lidar com seus orientandos; pela calma e paciência em suas orientações, mesmo que eu não tenha sido merecedor delas em alguns momentos; pela oportunidade, e pela grande confiança empregada desde nosso primeiro encontro. A você, desejo todo o bem, e espero poder continuar trabalhando ao seu lado.

Ao meu amigo e sócio, **Pindaro Cancian**, que me apresentou a oportunidade de desenvolver este projeto, e consequentemente ingressar no Mestrado de tecnologia em saúde.

A todos os docentes do **PPGTS**, pelo empenho e dedicação na transmissão de seus conhecimentos.

Ao **CNPq** pelo financiamento a este projeto de pesquisa e a **PUCPR** pela minha bolsa de Mestrado.

Aos colegas mestrandos, pela descontração e amizade, que foram essenciais em vários momentos.

Por fim, reservo este espaço para agradecer àquelas pessoas que mesmo sem saber, foram peças fundamentais para que eu hoje estivesse conquistando meu título de Mestre.

Agradeço a todos meus amigos, pela companhia e pelos inúmeros momentos de alegria, em especial, ao meu grande amigo **Diego**, que há mais de 10 anos me acompanha nas festas da vida.

Agradeço minha querida mãe, **Nadia**, e meu querido pai, **Paulo**, por todo o amor e carinho, e todo o empenho em me proporcionar uma vida digna e feliz. Espero sempre lhes dar orgulho e recompensá-los por tudo que já fizeram por mim.

Um enorme agradecimento a minha linda noiva **Ana Carolina**, cuja paciência em minhas ausências foi imprescindível para que eu chegasse até aqui, e cujo sorriso, amor, alegria e companhia diários me fazem tão bem. Deixo aqui registrado meu grande amor e admiração por você.

Enfim, nestes dois anos muitos se fizeram presentes e me ajudaram para que eu chegasse até aqui, fazendo este espaço ser insuficiente para citar e agradecer a todos, porém, fica aqui o meu muito obrigado e minha eterna gratidão a todos.

RESUMO

No Prontuário Eletrônico do Paciente são registradas informações sobre o atendimento ao paciente, algumas são armazenadas em formato de texto livre, e são chamadas de narrativas clínicas. Dentre as narrativas está o sumário de alta, que é uma síntese de informações essenciais, geradas durante a hospitalização do paciente, tornando-o uma ferramenta essencial para a transmissão de informações após a alta hospitalar. Um dos itens que deve estar presente nos sumários de alta são as informações relativas à continuidade do cuidado ao paciente, que, se não descritas de maneira adequada, podem implicar diretamente nos resultados assistenciais. Devido à importância dessa informação, uma ferramenta que alerte o profissional da saúde sobre a ausência desse tipo de informação pode impactar diretamente em uma melhora na assistência ao paciente. Portanto, o objetivo desta pesquisa foi elaborar um algoritmo para identificar informações relativas à continuidade do cuidado em meio aos textos dos sumários de alta. Os dados utilizados são provenientes do setor de cardiologia de um Hospital Universitário, contendo 5617 sumários de alta hospitalar. O projeto foi desenvolvido em cinco etapas principais. Inicialmente foi feita uma análise nos textos presentes no banco de dados. Os textos foram processados por algoritmos que implementam técnicas de Processamento de Linguagem Natural (PLN), e que já estão embutidas no software CoGrOO. Após processamento se fez necessário o uso de um *corpus* anotado da área médica, para obter melhores resultados. Além disso, um pré-processamento do texto foi essencial para reduzir erros provindos de sumários escritos de maneira errônea. O desenvolvimento do algoritmo de identificação de continuidade do cuidado, chamado de IRDischarge, foi a segunda etapa, que definiu o fluxo de processamento dos textos, e definiu-se que as ocorrências de informações seriam localizadas através de um banco de regras, baseado em padrões de etiquetas morfológicas, geradas pelo processamento dos textos através do CoGrOO. Para a terceira etapa, foram estabelecidos critérios de inclusão, exclusão e ordenação das regras e um banco de dados com 30 regras foi definido. A quarta etapa foi a avaliação de desempenho do algoritmo, que chegou a 93,3% de Precisão, 86,6% de Revocação e 91% de Especificidade. Como última etapa, foi desenvolvida uma ferramenta de apoio à escrita de sumários de alta, onde o algoritmo IRDischarge foi acoplado, para identificar informações de continuidade do cuidado, e alertar quanto a ausência desta importante informação durante a elaboração do mesmo. O algoritmo proposto permite a atualização do banco de regras, e permite utilização do mesmo para identificar outras informações de domínio clínico.

Palavras-chave: Sumários de alta do paciente. Processamento de Linguagem Natural, Continuidade do Cuidado ao Paciente.

ABSTRACT

The Electronic Health Records stores the information about the patient's treatment. Some of those information are not stored on a structured form, but like free-form text, and are called clinical narratives. Among the narratives, is the Discharge Summary, which is a summary of essential information, generated during patient's hospitalization, making of it an essential tool to transmit information after patient's discharge. Included in the items that must be present in the discharge summaries, are the information related to the patient's continuity of care that, if not described properly, results directly in patient's care results. Due to the importance of this information, a tool that alerts health professionals about the absence of such information may directly impact in improved patient care. For that reason, the objective of this research was to develop a method to identify information related to continuity of care among the texts of discharge summaries. The data used are from the cardiology sector of a University Hospital, containing 5617 discharge summaries. The project was developed in five main steps. Initially an analysis of the texts was made. The texts were processed by algorithms that implement the Natural Language Processing techniques (NLP), and are already built into the software CoGrOO. After processing the texts, the use of a corpus of medical domain was necessary for best results. In addition, a preprocessing of the text was essential to reduce clerical errors present in the discharge summaries. The development of the identification of the continuity of care algorithm, called IRDischarge was the second step. This step defines the processing flow of texts, and it was decided that the occurrences of information would be located through a database of rules, based on patterns of morphological tags, generated by the processing of texts by CoGrOO. For the third step, inclusion criteria, exclusion and ordering of rules were established. And a database with 30 rules was defined. The fourth step was to evaluate the performance of the method, which reached 93.3% Precision, Recall of 86.6% and Specificity of 91%. As a last step, was developed a tool to support the writing of discharge summaries, where the algorithm was incorporated, to identify information of patient's continuity of care, and alerting the absence of this important information during the preparation of the discharge summary. The proposed method allows upgrading the rules, and the use of it to identify other information of the clinical domain.

Keywords: Patient Discharge Summaries. Natural Language Processing, Continuity of Patient Care.

LISTA DE FIGURAS

Figura 1 - Técnicas mais frequentes no PLN	21
Figura 2 – Etapas da metodologia de desenvolvimento de regras.....	37
Figura 3 – Critérios para ordenação das regras	39
Figura 4 – Visão ampla dos processos de Geração das regras	42
Figura 5 - Fluxo de processos para identificação da continuidade do cuidado (IRDischarge)	47
Figura 6 – Gráfico comparativo de desempenho após primeiro processamento de cada iteração.....	52
Figura 7 – Ferramenta de apoio à escrita de sumários de alta com continuidade do cuidado.....	52
Figura 8 – Ferramenta de apoio à escrita de sumários de alta sem continuidade do cuidado.....	53

LISTA DE QUADROS

Quadro 1 – Exemplos de textos de sumários de alta	14
Quadro 2 - Proposta de padronização do conteúdo do sumário de alta	15
Quadro 3 - Exemplos de pesquisas que utilizam PLN em sumários de alta	22
Quadro 4 – Ferramentas de PLN avaliadas	26
Quadro 5 - Trechos que caracterizam continuidade do cuidado etiquetados.....	27
Quadro 6 - Significados de etiquetas morfológicas utilizadas em padrões de continuidade do cuidado	28
Quadro 7 - Comparativo de sumário de alta etiquetado, com e sem a fase pré-processamento.....	31
Quadro 8 – Exemplos de informações de retorno pós-alta	35
Quadro 9 – Exemplos de regras geradas na etapa de desenvolvimento inicial de regras	36
Quadro 10 – Trechos de sumários processados pelo IRDischarge e resultado do processamento.....	40
Quadro 11 – Exemplos de Especialização e Generalização de regras.....	41
Quadro 12 - Regras finais para identificação da continuidade do cuidado.....	48

LISTA DE TABELAS

Tabela 1 – Frequência de ocorrência de continuidade de cuidado por sentenças após a inversão na ordem das mesmas.....	34
Tabela 2 – Iterações executadas para geração de regras	43
Tabela 3 - Resultado do processamento dos sumários quanto às informações encontradas.....	50
Tabela 4 – Avaliação do processamento dos sumários de alta por iteração.....	51
Tabela 5 – Resultado final da avaliação do algoritmo IRDischarge	51

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
CEP	Comitê de Ética em Pesquisa
CID-10	Classificação Internacional de Doenças
CoGrOO	Corretor Gramatical acoplável ao LibreOffice
CONEP	Comissão Nacional de Ética em Pesquisa
FN	Falso Negativo
FP	Falso Positivo
GATE	<i>General Architecture for text engineering</i>
GT8	Grupo de Trabalho 8 de Arquitetura da Comissão de Estudos Especiais de Informática em Saúde da Associação Brasileira de Normas Técnicas
LAIS	Laboratório de Informática em Saúde
MedLEE	<i>A Medical Language Extraction and Encoding System</i>
NLTK	Natural Language Toolkit
PEP	Prontuário Eletrônico do Paciente
PLN	Processamento de Linguagem Natural
POS	<i>Part-of-speech</i>
PPGTS	Programa de Pós-graduação em Tecnologia em Saúde
PUCPR	Pontifícia Universidade Católica do Paraná
RI	Recuperação de Informação
VISL	<i>Visual Interactive Syntax Learning</i>
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO	8
1.1 CONTEXTUALIZAÇÃO DO PROJETO	10
1.2 HIPÓTESE DE TRABALHO	10
1.3 OBJETIVOS	10
1.3.1 Objetivo geral	10
1.3.2 Objetivos específicos	11
1.4 CONTRIBUIÇÕES CIENTÍFICAS	11
2 REVISÃO DE LITERATURA	12
2.1 NARRATIVAS CLÍNICAS	12
2.1.1 Sumários de alta	13
2.1.2 Continuidade de cuidado ao paciente	15
2.2 RECUPERAÇÃO DE INFORMAÇÃO EM SAÚDE	16
2.3 PROCESSAMENTO DE LINGUAGEM NATURAL	19
2.4 AVALIAÇÃO DE SISTEMAS DE RI	23
3 METODOLOGIA.....	25
3.1 ANÁLISE, ETIQUETAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS.....	26
3.1.1 Análise dos sumários de alta	28
3.1.2 Etiquetação	29
3.1.3 Pré-processamento	30
3.2 DESENVOLVIMENTO DO ALGORITMO DE IDENTIFICAÇÃO DE CONTINUIDADE DO CUIDADO, IRDISCHARGE.....	32
3.3 METODOLOGIA DE DESENVOLVIMENTO DE REGRAS	34
3.3.1 Etapa de desenvolvimento inicial de regras.....	35
3.3.2 Etapa de desenvolvimento iterativo de regras	37
3.4 AVALIAÇÃO DE DESEMPENHO DO ALGORITMO	43
3.5 DESENVOLVIMENTO DE FERRAMENTA PARA APOIO À ESCRITA DE SUMÁRIOS DE ALTA.....	44
4 RESULTADOS	46
4.1 ALGORITMO DE IDENTIFICAÇÃO DE CONTINUIDADE DO CUIDADO, IRDISCHARGE	46
4.2 BANCO DE REGRAS DO ALGORITMO.....	47
4.3 AVALIAÇÃO DE DESEMPENHO DO ALGORITMO	50
4.4 FERRAMENTA DE APOIO À ESCRITA DE SUMÁRIOS DE ALTA.....	52

5 DISCUSSÃO	54
5.1 ALGORITMO DE IDENTIFICAÇÃO DE CONTINUIDADE DO CUIDADO, IRDISCHARGE	54
5.2 BANCO DE REGRAS DO ALGORITMO	56
5.3 AVALIAÇÃO DE DESEMPENHO DO ALGORITMO	56
5.4 FERRAMENTA DE APOIO À ESCRITA DE SUMÁRIOS DE ALTA.....	57
5.5 TRABALHOS FUTUROS.....	58
6 CONCLUSÃO	60
REFERÊNCIAS.....	61

1 INTRODUÇÃO

O prontuário do paciente é um documento essencial para a assistência integral e continuada de um indivíduo. É um registro das informações retrospectivas, correntes e prospectivas, construído colaborativamente pela equipe multiprofissional de saúde que realiza o atendimento clínico. Engloba aspectos físicos, mentais e sociais do paciente (GALVÃO; RICARTE, 2011). Tem por objetivo assegurar serviços de saúde integrados de forma contínua, eficiente e com qualidade; auxiliar os profissionais responsáveis pelo cuidado ao paciente, proporcionando acesso às informações críticas ao atendimento (MASSAD; MARIN; NETO, 2003).

O Prontuário Eletrônico do Paciente (PEP) é a coleção de informações computadorizada relativa ao estado de saúde de um sujeito, armazenada e transmitida de forma segura, acessível aos usuários autorizados (ISO, 2005).

Existe diferenciação na forma em que os dados são registrados nos bancos de dados do PEP, podendo ser armazenados na forma de campos estruturados, como no caso dos dados, nome, data de nascimento e gênero; ou preenchidos na forma de texto livre, sendo estes denominados **narrativas clínicas**. As narrativas são textos utilizados para descrever de forma detalhada vários aspectos acerca do cuidado ao paciente (SHORTLIFFE; BARNETT, 2001; GALVÃO; RICARTE, 2011).

O **sumário de alta** é uma das narrativas clínicas que pode estar contida no prontuário do paciente, e pode ser descrito como uma síntese dos acontecimentos clínicos (sintomas, diagnósticos investigados, procedimentos realizados, medicamentos utilizados, intercorrências e outros), ocorridos durante o atendimento. Tem como objetivo principal auxiliar na transição do cuidado ao paciente, para que ela ocorra de forma mais segura, realizando a comunicação entre os profissionais responsáveis pelo cuidado ao paciente e atenção primária. O sumário de alta é também utilizado para elucidar o paciente sobre o autocuidado pós-atendimento (PINTO, 2006).

Apesar de já existir uma proposta de padronização do conteúdo que um sumário de alta deve conter, desenvolvida por Andréia Cristina de Souza, juntamente com o grupo GT8 de Arquitetura da Comissão de Estudos Especiais de Informática em Saúde da ABNT, em sua maioria, os sumários ainda não são escritos seguindo o padrão definido, apresentando a ausência parcial e/ou completa de dados essenciais ao paciente (SOUZA, 2012).

Outra característica presente na escrita dos sumários é a forma não estruturada na qual eles são definidos, ou seja, contendo diversos dados de cunho descritivo junto a dados pessoais e outros dados já localizados em algum outro campo estruturado na base de dados do PEP (UZUNER *et al.*, 2008).

Sendo muitas das informações contidas no sumário de alta vitais ao tratamento, é importante que sistemas de informação tenham a habilidade de recuperá-las, não só para auxílio ao cuidado e apoio à decisão, mas também para evidenciar a incompletude ou até mesmo ausência dessas informações fundamentais nos textos escritos pelos profissionais da saúde (KRIPALANI *et al.*, 2007).

Há mais de duas décadas pesquisadores vêm desenvolvendo métodos para **Recuperação de Informação** (RI) em narrativas clínicas, fazendo uso, na maioria dos casos, das técnicas de **Processamento de Linguagem Natural** (PLN) (XU *et al.*, 2010). Tendo como maior desafio lidar com textos com alta frequência de erros e com pouca estrutura formal definida (ERTMER; ÜCKERT, 2005).

Entre os trabalhos encontrados na literatura, que utilizam PLN em narrativas clínicas ou na recuperação de informações relacionadas à saúde, destacam-se pelos bons resultados na recuperação de informações e similaridade com este projeto:

- a) MedLEE (Medical Language Extraction and Encoding System), desenvolvido por Friedman e sua equipe de pesquisa (FRIEDMAN *et al.*, 1994), que fora inicialmente criado para apoio à decisão baseado em relatórios radiológicos, foi estendido para outros domínios, como os sumários de alta (FRIEDMAN *et al.*, 1999);
- b) MedEx (A Medication Information Extraction System for Clinical Narratives), foi o resultado do trabalho de pesquisadores do Departamento de Informática Biomédica da Universidade de Vanderbilt, que consiste na extração de informações relacionadas a medicamentos de sumário de alta (XU *et al.*, 2010);
- c) Bui e seus colaboradores especificaram um método/ferramenta para extrair e combinar relacionamentos entre medicamentos para tratamento do HIV e mutações nos genomas virais atrelados a doença encontrados na literatura (BUI *et al.*, 2010);
- d) e por fim, a dissertação de Hugo Bulegon, que descreve um método

baseado em regras para identificação de doenças contidas em sumários de alta, e fazendo posterior mapeamento para classificação internacional de doenças (BULEGON; MORO; GIRARDI, 2010; BULEGON, 2011).

Nesta pesquisa, além da identificação e recuperação de informações do texto livre descrito nos sumários de alta, também está presente no escopo a evidênciação das informações faltantes nos sumários de alta, com o intuito de alertar o profissional da saúde, vislumbrando a escrita de textos de alto nível de completude e que resultem em maior qualidade dos textos, visando uma melhoria na transição do cuidado ao paciente.

A partir deste contexto, o objetivo desta pesquisa é a identificação da presença de informações relacionadas à continuidade do cuidado ao paciente, nos sumários de alta escritos na língua portuguesa. Para tal, desenvolveu-se um algoritmo para identificação das informações, bem como, uma ferramenta de auxílio à escrita desta narrativa clínica, que alerte o profissional sobre a ausência de informações referentes à continuidade do cuidado.

1.1 CONTEXTUALIZAÇÃO DO PROJETO

Esta pesquisa é parte do projeto “Recuperação Semântica de Informações em Bases Multilíngue Visando o Apoio à Decisão na Área da Saúde”, com fomento do CNPq sob Edital 009/2010, realizada entre janeiro de 2011 e janeiro de 2014.

1.2 HIPÓTESE DE TRABALHO

Visando uma melhoria na qualidade do conteúdo dos textos das narrativas clínicas, em especial dos sumários de alta, esta pesquisa busca demonstrar que a partir da utilização de técnicas de PLN em textos não estruturados de sumários de alta é possível identificar a presença ou ausência de informações relativas à continuidade do cuidado ao paciente.

1.3 OBJETIVOS

1.3.1 Objetivo geral

O objetivo desta pesquisa é identificar nos textos de sumários de alta hospitalar a presença de informações referentes à continuidade do cuidado ao

paciente.

1.3.2 Objetivos específicos

- a) Elaborar um algoritmo para identificar as informações referentes à continuidade do cuidado ao paciente em sumários de alta.
- b) Avaliar o desempenho do algoritmo de identificação de continuidade do cuidado.
- c) Desenvolver uma ferramenta computacional baseada no algoritmo proposto para auxiliar na elaboração dos sumários de alta que indique a ausência de continuidade do cuidado ao paciente.

1.4 CONTRIBUIÇÕES CIENTÍFICAS

As contribuições científicas deste projeto de pesquisa residem no algoritmo baseado em regras morfológicas e em técnicas de PLN para a identificação da presença de informações relativas à continuidade do cuidado, nas próprias regras geradas para este fim, e principalmente, no método elaborado para geração e melhoria de regras, que especifica critérios para geração de um novo banco de regras ou aperfeiçoamento de um pré-existente.

2 REVISÃO DE LITERATURA

A revisão de literatura foi realizada com o propósito de investigar as informações referentes à continuidade do cuidado ao paciente contidas em sumários de alta, bem como estudar metodologias utilizadas na recuperação de informações em narrativas clínicas, e então, estabelecer as áreas de estudo e técnicas envolvidas na elaboração do algoritmo.

Os trabalhos de pesquisa que referenciaram esta revisão foram selecionados em diversas bases de dados científicas, por meio de filtragem de palavras-chave encontradas no *abstract* e leitura do *abstract* dos mesmos. Optou-se pela escolha de artigos publicados nos últimos dez anos, contados a partir do início de desenvolvimento do projeto (2003-2013), os que estão fora deste intervalo de tempo ou não são encontrados pela busca às bases de dados científicas, são justificados por serem considerados textos clássicos no assunto, e/ou trazem definições essenciais para a fundamentação de alguma das áreas presentes no estudo.

As bases de dados científicas pesquisadas foram: PubMed, IEEEExplore, ScienceDirect e Google Scholar. As entradas das pesquisas foram:

- a) (“Information Retrieval” OR “Natural Language Processing”) AND “Clinical Narratives”
- b) (“Information Retrieval” OR “Natural Language Processing”) AND “Discharge summaries”
- c) “Continuity of care” AND “Clinical Narratives”

2.1 NARRATIVAS CLÍNICAS

Shortliffe e Barnett definem os dados médicos em três grandes grupos: os numéricos e mensuráveis, como temperatura, pulso, pressão e alguns exames, por exemplo; as imagens e gráficos; e os narrativos, entre eles: a descrição de sintomas do paciente, histórico assistencial, entre outros. (SHORTLIFFE; BARNETT, 2001)

Dentre os dados armazenados no Prontuário Eletrônico do Paciente, constam informações dispostas em forma de texto livre, entre elas: laudos cardiológicos, sumários de alta, evoluções, anamneses, entre outras. Estes documentos são chamados de narrativas clínicas (PACHECO, 2009).

As narrativas clínicas compõem de forma essencial o Prontuário do Paciente.

O texto contido nesses documentos é de extrema importância no que se refere ao atendimento ao paciente, devido à riqueza de informações acerca dos achados clínicos do paciente (CHEN; HRIPCSAK; FRIEDMAN, 2006). Elas relatam a história da assistência ao paciente e a perspectiva médica frente ao panorama geral da situação clínica atual e plano de cuidado, tornando-se a fonte mais completa de análise das informações relacionadas ao cuidado do paciente (SHEN *et al.*, 2012).

Grande parte dos dados clínicos disponíveis encontram-se em forma de narrativa, como resultado da transcrição de ditados, escrita direta pelo profissional da saúde ou até mesmo pelo uso de sistema de reconhecimento de fala. O formato de texto livre, característico das narrativas clínicas, é conveniente para expressar conceitos e/ou eventos advindos da assistência, porém, é um fator dificultador para a pesquisa, sumarização, apoio à decisão ou análises estatísticas das informações (MEYSTRE *et al.*, 2008).

2.1.1 Sumários de alta

Em meio às narrativas clínicas armazenadas no PEP, encontra-se o sumário de alta, também chamado de resumo de alta, que é o principal documento para se gravar e recuperar as informações geradas durante a hospitalização do paciente, tais como: intervenções diagnósticas e clínicas, relatórios de evolução e orientações acerca da continuidade do cuidado (MYERS *et al.*, 2006; LAPELLE *et al.*, 2006).

Por ser composto de informações importantes para o cuidado ao paciente, o sumário de alta se torna uma ferramenta essencial para a transmissão de informações relevantes após a alta hospitalar ou até mesmo uma transferência entre departamentos do hospital. Dados relacionados às prescrições médicas, diagnósticos e encaminhamentos são aspectos fundamentais, pois as ações dos médicos posteriormente envolvidos no tratamento serão baseadas nos mesmos. (PERREN *et al.*, 2008). Há casos em que o sumário de alta é o único meio de comunicação entre o hospital e o médico de cuidado primário (VAN WALRAVEN; SETH; LAUPACIS, 2002).

Devido ao formato livre de escrita, os sumários de alta estão suscetíveis a uma série de problemas, como ausência de informações e erros na entrada dos dados, que podem afetar na qualidade da assistência, bem como, da continuidade do cuidado ao paciente (VAN WALRAVEN; WEINBERG, 1995; VAN WALRAVEN;

ROKOSH, 1999). A seguir, exemplos de textos de sumários de alta, com falta de informações, erros gramaticais, ou escrita completamente em letras maiúsculas. (Quadro 1).

Quadro 1 – Exemplos de textos de sumários de alta

Textos de sumários de alta
Paciente hipertenso, portador de intolerância a carboidratos (sic) e cardiopatia isquêmica, CAT mostrando lesões em DAE, DPCD e MgCX, internou eletivamente para angioplastia. Realizada ACTP com implante de Stent em MgCX, com sucesso angiográfico e sem intercorrências.
PACIENTE CARDIOPATA ISQUÊMICO GRAVE COM INSUFICIÊNCIA CARDÍACA CLASSE III-IV, JÁ REALIZOU DUAS CIRURGIAS DE REVASCULARIZAÇÃO DO MIOCÁRDIO PRÉVIAS, POSTERIORMENTE APRESENTANDO OCLUSÃO DAS PONTES, INTERNOU PARA REALIZAR ACTP+STENT DE TROCO DE CORONÁRIA E. PROCEDIMENTO REALIZADO SEM INTERCORRÊNCIAS COM BOM RESULTADO ANGIOGRÁFICO. CERCA DE 4H APÓS O PROCEDIMENTO APRESENTOU PCR EM FIBRILAÇÃO VENTRICULAR SENDO REANIMADO COM SUCESSO(NOVO CAT SEM CMPLICAÇÕES MECÂNICAS EM STENT IMPLANTADO). TRANSFERIDO PARA CTI ONDE FOI MANEJADO COM BALÃO INTRA-AÓRTICO + S-GANZ+ VENTILAÇÃO MECÂNICA, APRESENTOU BOA EVOLUÇÃO INICIALMENTO SENDO RETIRADO DA V. MEC.. EVOLUIU COM INSUFICIÊNCIA RESPIATÓRIA POR INFECÇÃO RESPIRATÓRIA ASSOCIADO À CHOQUE SÉPTICO EVULUINDO PARA ÓBITO MESMO COM OS CUIDADOS INTENSIVOS INSTITUIDOS.
Masculino, 61 anos, infarto do miocárdio em 2003, CRM em março de 2003 (marginal para DAE, safenas para primeira e segunda marginais e descendente posterior), reiniciou há 3 meses com dor torácica. Cateterismo cardíaco em fevereiro de 2004 mostra safenas para primeira e segunda marginais ocluídas e lesão grave proximal na safena para descendente posterior. Realizado sem intercorrências no dia 27 de fevereiro de 2004 angioplastia coronária percutânea+implante de stent em safena para descendente posterior. Apresentou boa evolução, recebendo alta assintomático, sinais vitais estáveis, exame físico dos membros inferiores normais, com as seguintes orientações: - retorno à emergência se necessário - retorno ao seu médico assistente - uso das seguintes medicações: --- sinvalip 80mg - meio cp via oral à noite --- selozok 50mg - 1 cp via oral 1xdia --- captopril 25mg - 1 cp via oral 12/12 horas --- AAS 100mg - 1 cp via oral no almoço --- ticlopidina 250mg - 1 cp via oral 12/12 horas

Kripalani e colaboradores afirmam que para um sumário de alta ter qualidade, as informações devem ser descritas de forma organizada, apresentando dados essenciais de forma objetiva, para facilitar o entendimento por parte do profissional que dará continuidade na assistência (KRIPALANI *et al.*, 2007; LONG, 2005).

Estudos comprovam que com o advento do PEP, no qual a documentação

médica é escrita com a utilização de uma ferramenta computacional, houve um aumento no tamanho dos textos escritos, redundância, má formatação e desordem na apresentação das informações (EMBI *et al.* 2004).

No Brasil, existe uma iniciativa para formalização e padronização dos sumários de alta, elaborada por Andréia Cristina de Souza (SOUZA, 2012), e adotada pelo GT8 - Grupo de Arquitetura da Comissão de Estudos Especiais de Informática em Saúde da ABNT. Na pesquisa de Souza, foram definidas as informações que devem ser apresentadas no conteúdo do texto e sua ordem de aparição no texto, contendo 9 elementos-chave, com cada um deles apresentando um conjunto de itens que devem ser observados durante a elaboração do sumário de alta. A proposta de padronização está resumida no Quadro 2.

Quadro 2 - Proposta de padronização do conteúdo do sumário de alta

Ordem	Elementos-chave
1	Identificação do paciente e caracterização da internação
2	Motivo da admissão, diagnósticos e comorbidades
3	Achados clínicos relevantes e outros achados
4	Procedimentos diagnósticos e terapêuticos realizados
5	Medicações importantes, inclusive as de alta
6	Exames pendentes
7	Evolução e condições de estado do paciente na alta
8	Instruções relativas ao acompanhamento
9	Destino da alta e ou retorno / Anexos

Fonte: Adaptado de Souza, 2012

2.1.2 Continuidade de cuidado ao paciente

A continuidade do cuidado ao paciente, entre as várias definições encontradas na literatura, é apresentada por Haggerty e colaboradores como uma gestão multidisciplinar de cuidado a saúde do paciente ao longo do tempo (HAGGERTY *et al.*, 2003).

Há décadas estudos vêm sendo feitos para comprovar que a boa comunicação entre os envolvidos no tratamento ao paciente é fundamental para uma continuidade do cuidado de qualidade. Em 1980, Gonnella e Herman já estudavam o impacto de uma boa transição do cuidado nos resultados do tratamento, satisfação dos pacientes e nos custos do cuidado (GONNELA; HERMAN, 1980).

O principal canal de comunicação entre as partes envolvidas na assistência ao paciente é o sumário de alta. Segundo Kergoat e seus colaboradores, a falta de informação na descrição da continuidade do cuidado no sumário de alta implica diretamente nos resultados assistenciais, que vão desde duplicação desnecessária de consultas e investigações médicas, até um agravamento do estado de saúde e, conseqüentemente, insatisfação do paciente (KERGOAT *et al.*, 2010).

Moore e colaboradores afirmam que aproximadamente metade dos pacientes internados vão vivenciar pelo menos um erro médico relacionado à continuidade de medicações ou cuidados de acompanhamento (MOORE *et al.*, 2003). A evidenciação da importância de uma melhor comunicação entre os prestadores de cuidados de saúde para diminuição de erros de medicações e cuidados a saúde pós-alta hospitalar é apresentada na pesquisa de Roughead e colaboradores (ROUGHREAD *et al.*, 2009).

O maior envolvimento dos profissionais da saúde dos hospitais no atendimento à pacientes internados, que anteriormente era apenas tratado pelo profissional de atendimento primário, introduz a descontinuidade ao tratamento, que pode ser amenizada com uma melhor comunicação e documentação, especialmente com sumários de alta de alta qualidade. Stetson e colaboradores desenvolveram uma ferramenta para avaliar a qualidade dos sumários de alta com o objetivo de apoiar as transições de pacientes e contribuir para maior segurança do paciente (STETSON *et al.* 2008).

Na pesquisa de Souza (2012), onde foi proposta a padronização dos sumários de alta, as informações referentes à continuidade do cuidado, exceto no que diz respeito à parte medicamentosa, estão definidas no elemento-chave de número 9, apresentado como Destino da alta e ou retorno (conforme Quadro 2). As informações previstas para esta seção do sumário de alta são: orientações de continuidade de tratamento, encaminhamento para outro profissional ou especialidade, previsão de retorno pós-alta, local para acompanhamento do paciente, exames e possíveis transferências (SOUZA, 2012).

2.2 RECUPERAÇÃO DE INFORMAÇÃO EM SAÚDE

Recuperação de Informação (RI) é o nome dado à área de estudo que trata da representação, armazenamento, organização e acesso a informações em

documentos, na *Web*, registros estruturados e semi-estruturados, de tal modo que, permita aos usuários acesso fácil às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO; 2011).

Um dos maiores desafios, quando se trata de RI em saúde, é a identificação/extração das informações contidas em narrativas clínicas, e a utilização das mesmas em conjunto com informações vindas de aplicações de gerenciamento de pacientes, apoio à decisão, e pesquisa clínica. Para identificação/extração dessas informações, sistemas que se utilizam de PLN são utilizados (CHEN; HRIPCSAK; FRIEDMAN, 2006).

A rápida adoção do PEP em ambientes hospitalares, em paralelo com o crescimento da quantidade de dados, em forma de narrativas, armazenados eletronicamente, juntamente com as necessidades de melhoria na qualidade de atendimento e redução dos erros médicos, são fatores determinantes para o desenvolvimento de sistemas de RI em narrativas clínicas (MEYSTRE *et al.*, 2008).

Como as narrativas podem ser relacionadas a diversas áreas de atenção à saúde (radiologia, cardiologia, neurologia, etc.) e armazenam as mais variadas informações (intervenções cirúrgicas, informação medicamentosa, diagnósticos, etc.), os métodos propostos geralmente são desenvolvidos para domínios específicos, e em alguns casos, adaptados em pesquisas futuras para busca em outros tipos de documentação e diferentes dados colocados como objetivo da RI.

A pesquisa de Friedman e colaboradores (1994) foi um dos marcos na área e teve sua metodologia adaptada para diferentes situações ao longo do tempo. Inicialmente, o método tinha como proposta transformar as informações relacionadas à radiologia em um vocabulário de termos controlado, tendo como objetivo o auxílio à decisão e à pesquisa neste domínio. Um dos diferenciais do algoritmo é que ele se utilizou de elementos de diferentes metodologias de RI combinadas, como a comparação baseada em padrões de texto, baseada no valor semântico de palavras e também no valor sintático de termos. Além disso, o sistema, que posteriormente foi nomeado como MedLEE (FRIEDMAN *et al.*, 1995), foi construído de maneira extensível, permitindo a adição de novos domínios de uso na saúde e funcionalidades, de forma modular e sistemática, sem afetar a abordagem inicialmente desenvolvida (FRIEDMAN *et al.*, 1994; FRIEDMAN *et al.*, 1999; GUNDLAPALLI *et al.*, 2008; WANG *et al.*, 2009; CHIANG; LIN; YANG, 2010; WU *et al.*, 2012).

Xu e seus colaboradores tinham a necessidade de extrair informações detalhadas acerca de medicamentos presentes em narrativas clínicas. Vários sistemas desenvolvidos, incluindo o MedLEE, já extraem de forma precisa nomes de medicamentos, porém, durante a pesquisa percebeu-se que os dados que compõem as informações medicamentosas, como a dosagem, duração, frequência de uso, entre outros, não tem as altas taxas de precisão encontradas na identificação do nome do medicamento. Assim, foi proposto um novo método (MedEx) para a identificação completa dos dados de medicamentos. Por intermédio da análise manual de sumários de alta utilizados na pesquisa, foi gerado um modelo de informações para ser inserido no conjunto de treinamento do algoritmo e regras específicas elaboradas para cada padrão de definição de medicamento encontrado. O algoritmo faz a leitura do texto, e envia-o para um etiquetador semântico, que baseado nos modelos inseridos previamente, assinala o valor dos termos dentro das categorias estabelecidas. Com os termos definidos em suas categorias, são aplicadas as regras definidas para a separação e identificação de cada um dos dados que fazem parte da definição do medicamento. Como resultado o algoritmo teve taxas de precisão superiores a 95% na identificação de nove dos onze dados pretendidos (XU *et al.*, 2010).

Em 2010, um método para extrair e combinar relacionamentos entre mutações em genomas virais e medicamentos para o HIV foi desenvolvido por Bui e sua equipe de pesquisa. Diferentemente dos métodos já citados, a origem dos dados usados nesta pesquisa não é provinda de narrativas clínicas, e sim, de *abstracts* de artigos científicos buscados na base de dados do PubMed. O método foi definido em três fases:

- a) recuperação de texto - a primeira fase consiste em obter os *abstracts* no PubMed utilizando como filtro uma série de nomes de medicamentos usados no tratamento ao HIV e a partir deles selecionar através do uso de expressões regulares, sentenças que contenham relação com mutações;
- b) pré-processamento de texto - nesta fase, são executados vários processos para a simplificação dos textos, entre eles: substituição de termos conhecidos por suas formas abreviadas, normalização das sentenças e agrupamento de mutações e nomes de medicamentos. Após esses processos, são geradas as relações gramaticais dos termos e com elas as regras de relação que serão utilizadas na última fase;

c) extração de relacionamentos - a partir das regras, as relações entre mutações e medicamentos são extraídas e combinadas pelo algoritmo.

A avaliação dos resultados mostrou que o método teve 87% de precisão na identificação de relações entre mutações e medicamentos (BUI *et al.*, 2010).

Bulegon desenvolveu um método para identificação de diagnósticos contidos em sumários de alta e mapeamento para o CID-10. Utilizando técnicas de PLN, os sumários de alta foram etiquetados morfológicamente e analisados manualmente na busca de relações entre os diagnósticos e suas etiquetas. A partir dessa análise, regras com três termos (etiquetas) foram definidas, sempre montadas na estrutura: [antecessor][diagnóstico][sucessor]. Para avaliação dos resultados o algoritmo fez a leitura de sumários de alta na tentativa de encontrar os padrões descritos no banco de regras em meio ao texto. Uma vez encontrado, efetuou-se a correlação do diagnóstico encontrado para o CID-10 (BULEGON, 2011).

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural é uma área de pesquisa que atua no entendimento e manipulação de texto escrito em linguagem natural, dispondo de ferramentas e técnicas para o desenvolvimento de aplicações computacionais. O PLN caracteriza-se pela multidisciplinaridade, pois, envolve várias áreas de pesquisa, entre elas: Inteligência Artificial e Linguística (CHOWDHURY, 2003).

Existem duas principais motivações para o desenvolvimento da área de PLN, a científica e a tecnológica, sendo a científica para o entendimento completo da natureza da linguagem. A motivação tecnológica explica-se pela necessidade da melhoria na comunicação entre homens e máquina (MANARIS, 1998).

Inicialmente, o PLN era algo distinto da RI em textos, porém, com o tempo, o PLN e a RI convergiram em alguns pontos. O histórico da evolução do PLN, bem como uma introdução às técnicas mais utilizadas e dificuldades encontradas na utilização dos métodos em textos clínicos são abordados no trabalho feito por Nadkarni e seus colaboradores (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Os conhecimentos necessários para desenvolver aplicações complexas que envolvam linguagem natural são definidos em seis categorias/etapas (JURAFSKY; MARTIN, 2008):

- a) Fonética e Fonologia – que estuda os sons provenientes da linguagem;
- b) Morfologia – que estuda as categorias de classificação das palavras;
- c) Sintaxe – que estuda os relacionamentos estruturais entre as palavras;
- d) Semântica – que estuda o significado das palavras;
- e) Pragmática – que estuda a junção das outras etapas para apresentação do resultado final;
- f) Discurso – que estuda unidades linguísticas maiores que uma única expressão.

Nem todas as etapas descritas participam de todas as aplicações que se utilizam de PLN, a fonética, por exemplo, é utilizada geralmente quando se tratam de textos registrados em áudio apenas. Na Figura 1 é apresentado um diagrama que ilustra as técnicas de PLN mais comuns em sistemas de RI. Na figura as técnicas são definidas em inglês e português, nesta pesquisa as técnicas serão nomeadas em inglês por serem amplamente usadas desta forma, inclusive em trabalhos acadêmicos e pesquisas escritas em português.

Os primeiros processos executados quando se trata de PLN são, a delimitação de sentenças e a separação dos *tokens* (que são os termos e pontuações encontrados em meio ao texto) (PACHECO, 2009).

Quando informações linguísticas acerca dos termos são necessárias, um etiquetador gramatical (*part-of-speech tagger*) identifica, com a colocação de uma etiqueta (*tag*), a categoria gramatical de cada palavra do texto. Este etiquetador geralmente é morfológico (identifica somente a categoria morfológica da palavra) ou morfo-sintático (identifica também as funções sintáticas) (GONZALEZ; LIMA; LIMA, 2006).

Os etiquetadores têm dois modelos fundamentais: o baseado em regras e o estatístico. No primeiro, são geradas regras para definição de cada categoria existente, para assim fazer a etiquetação do item verificado. Já o segundo modelo utiliza-se do treinamento de um *corpus* previamente anotado, com termos e suas categorias correspondentes de etiqueta associados a cada palavra, assim, para a definição da categoria de um item, são feitos cálculos de probabilidade para determinar a categoria na qual ele será etiquetado. (JURAFSKY; MARTIN, 2008).

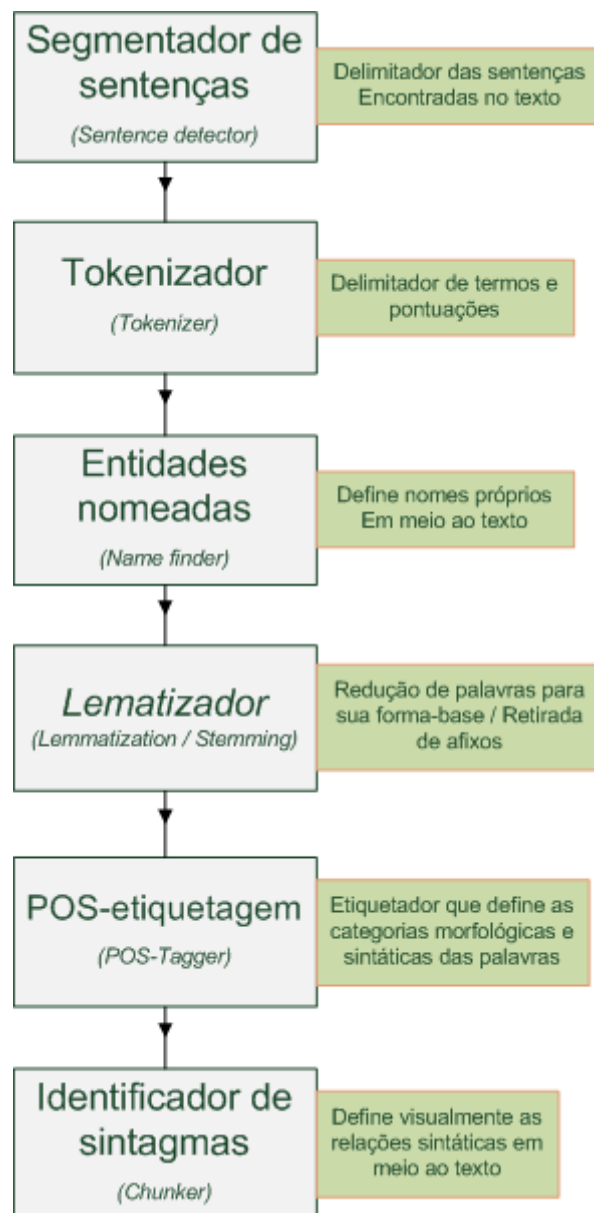


Figura 1 - Técnicas mais frequentes no PLN

Na área de RI e PLN, a definição de *corpus* é dada a uma coleção de textos utilizada como base de consulta e comparação estatística para uma futura etiquetagem de textos. Um *corpus* é dito como “anotado” quando um especialista marca as palavras contidas nos textos com suas categorias morfológicas ou relações sintáticas, para melhor aproveitamento de um etiquetador estatístico no domínio de aplicação desejado (saúde, financeiro, jornalístico, etc.) (MANNING; SCHUTZE, 1999).

O dicionário que armazena as categorias morfológicas possíveis é conhecido como léxico ou *tagset*. O PLN ainda pode usar processos de inferência para a etiquetagem de palavras desconhecidas pelo *corpus*, e também utilizar-se de regras

pré-estabelecidas para gerar informações complementares de etiqueta, como tempo verbal, gênero (masculino, feminino) e número (singular, plural) de uma palavra (MANNING; SCHUTZE, 1999; MENEZES; NETO, 2002).

Em muitos casos é interessante ao sistema de PLN normalizar o texto, com o objetivo de buscar palavras com sentido semelhante. A normalização linguística é dividida em três áreas: morfológica, sintática e léxico-semântica (ARAMPATZIS *et al.*, 1999).

A normalização morfológica ocorre quando há redução lexical (na estrutura da palavra), buscando combinar a representação de dois ou mais termos em um único conceito (ou significado), reduzindo os termos a uma única forma conceitual. Os processos de equivalência mais comuns são o *Stemming* e *Lemmatization* (BITENCOURT, 2006).

O *Stemming*, através de um processo heurístico de eliminação de afixos de derivação ou flexão, reduz uma palavra a sua parte semelhante ao seu radical (*stem*) (ORENGO; HUYCK, 2001). Por exemplo, o *Stemming* das palavras “ferro”, “ferrar” e “ferrado”, resultam em um mesmo termo: “ferr”.

O *Lemmatization* através do uso de um vocabulário altera os verbos para sua forma infinitiva, e substantivos e adjetivos à forma masculina e singular da palavra. (ARAMPATZIS *et al.*, 1999). Por exemplo, o *Lemmatization* das palavras “seguimento” e “seguir”, resultam em um mesmo termo: “seguir”.

O PLN tem a capacidade de identificar e extrair informações em narrativas clínicas através de suas técnicas, como várias pesquisas o vem fazendo. O Quadro 3 apresenta exemplos de pesquisas que se utilizam de PLN para recuperação de informações em sumários de alta.

Quadro 3 - Exemplos de pesquisas que utilizam PLN em sumários de alta

Autor	Ano	Informações identificadas
Genevieve <i>et al.</i>	2005	Eventos adversos
Zeng <i>et al.</i>	2006	Diagnósticos, co-morbidades e situação tabagista
Xu <i>et al.</i>	2010	Informações de medicamentos
Bui <i>et al.</i>	2010	Informações de medicamentos e mutações do vírus HIV
Matheny <i>et al.</i>	2012	Sintomas de infecção

2.4 AVALIAÇÃO DE SISTEMAS DE RI

Para a avaliação de sistemas de RI e mensuração de sua efetividade alguns métodos foram definidos. Geralmente a eficiência do algoritmo é medida utilizando os números finais de informações recuperadas pelo sistema, sendo as possibilidades definidas como.

- verdadeiros positivos (VP) – foi recuperada alguma informação e ela pertence ao grupo de informações desejadas;
- falsos positivos (FP) – foi recuperada alguma informação, porém, ela não pertence ao grupo de informações desejadas;
- verdadeiros negativos (VN) – não foi recuperada nenhuma informação e o documento não continha nenhuma informação pertencente ao grupo de informações desejadas;
- falsos negativos (FN) – não foi recuperada nenhuma informação, porém, o documento contém informação pertencente ao grupo de informações desejadas.

A partir desses valores, é possível calcular através de fórmulas algumas características do sistema e avaliá-lo (MANNING; RAGHAVAN; SCHÜTZE, 2008).

Para cálculo da **Precisão** (*Precision*) do sistema, que é a proporção de informações selecionadas pelo sistema de maneira correta (quantas respostas obtidas que estão corretas), é utilizada a seguinte fórmula:

$$\textbf{Precisão} = \frac{VP}{VP+FP} \quad (1)$$

Para cálculo da **Revocação** (*Recall*) ou **Sensibilidade** (*Sensitivity*) do sistema, que é a proporção de informações desejadas que o sistema selecionou corretamente (quantas respostas corretas o sistema obteve), é utilizada a seguinte fórmula:

$$\textbf{Revocação} = \frac{VP}{VP+FN} \quad (2)$$

A **Especificidade** (*Specificity*) do algoritmo, que é a habilidade do sistema identificar resultados negativos. Para este cálculo, emprega-se a equação 3:

$$\textit{Especificidade} = \frac{VN}{VN+FP} \quad (3)$$

Quando se trata de sistemas de RI, onde o mesmo pode ter 100% de Revocação, porém, uma taxa muito baixa de Precisão, é conveniente combinar estas duas características em um cálculo único. Para tal, existe a possibilidade de se fazer uma média harmônica dos dois, chamada ***F-Measure***, calculada conforme fórmula a seguir:

$$\textit{F - Measure} = 2 * \frac{\textit{Precisão*Revocação}}{\textit{Precisão+Revocação}} \quad (4)$$

3 METODOLOGIA

O presente trabalho de pesquisa é definido como um estudo experimental, com característica exploratória, descritiva e de desenvolvimento. Os resultados foram avaliados de forma quantitativa, utilizando-se de metodologias de avaliação utilizadas em sistemas de RI.

A etapa inicial de desenvolvimento da pesquisa foram os testes para a identificação da continuidade do cuidado por meio da replicação do método de Bulegon (2011), que a partir dos trabalhos encontrados na Revisão de Literatura mostrou-se mais adequado ao objetivo desejado neste projeto, pois além de trabalhar com sumários de alta, inclusive com o mesmo conjunto de dados, também se refere à identificação de um conteúdo presente nos textos de sumários de alta.

Após o uso do método de Bulegon, existiam ainda dúvidas quanto à ferramenta de PLN a ser utilizada. O uso de outras ferramentas foi avaliado (conforme Quadro 4), e ficou evidenciado que a melhor opção seria continuar utilizando o CoGrOO, como no método de Bulegon, pois ele implementa uma funcionalidade chamada *Featurizer*, que aplica juntamente a etiqueta morfológica das palavras, também as informações de tempo verbal, gênero e número das palavras em português, permitindo que o método se valha de regras mais específicas na busca de informações. Além disso, o uso do *corpus* médico já anotado, construído por Peters e seus colaboradores, em 2010, poderia ser usado novamente na pesquisa.

Outro fator que determinou a escolha do CoGrOO foi que ele se utiliza da linguagem de programação Java, que é mais difundida atualmente que a linguagem Python, a partir da qual o NLTK (Natural Language Toolkit) é baseado.

A versão do CoGrOO utilizada foi a 3.0.5. Houve uma tentativa de utilizar a nova versão 4, porém, além dela estar em versão *beta*, o *tagset* utilizado foi alterado, e para que fosse possível a utilização do *corpus* médico disponível, ele teria de ser reanotado a partir desse novo *tagset*. Optou-se, então, por continuar com a versão 3.0.5, pois as alterações da nova versão não são necessárias neste trabalho.

Quadro 4 – Ferramentas de PLN avaliadas

Ferramenta	Possibilidade de inserção de corpus anotado	Etiquetagem morfológica via corpus	Inferência de etiquetagem (gênero, número, etc.)	Linguagem de programação	Regras para a língua portuguesa
CoGrOO	Sim	Sim	Sim	Java	Sim
OpenNLP	Sim	Sim	Não	Java	Não
GATE	Sim	Sim	Não	Java	Não
NLTK	Sim	Sim	Sim	Python	Não

Os textos dos sumários de alta utilizados nesta pesquisa foram elaborados durante atendimentos do Serviço de Cardiologia de um Hospital Universitário, realizados de junho de 2002 a maio de 2007 e todas as altas de julho de 2007, totalizando 5.617 sumários de alta. O uso da base de dados com os sumários de alta foi aprovada por meio do protocolo da Comissão Nacional de Ética em Pesquisa (CONEP) 0375.084.000-10, tendo o parecer favorável do Comitê de Ética em Pesquisa (CEP) da PUCPR nº. 0004422/10.

O trabalho foi definido em cinco etapas principais:

- 1) Análise, etiquetagem e pré-processamento dos dados;
- 2) Desenvolvimento do algoritmo de identificação de continuidade do cuidado, *IRDischarge*;
- 3) Desenvolvimento do banco de regras do algoritmo e de uma metodologia de geração de regras;
- 4) Avaliação de desempenho do algoritmo;
- 5) Desenvolvimento de ferramenta para apoio à escrita de sumários de alta.

O local de desenvolvimento desta pesquisa foi o Laboratório de Informática em Saúde (LAIS) do Programa de Pós-Graduação de Tecnologia em Saúde (PPGTS) da Pontifícia Universidade Católica do Paraná (PUCPR).

3.1 ANÁLISE, ETIQUETAGEM E PRÉ-PROCESSAMENTO DOS DADOS

O algoritmo para identificação dos trechos referentes à continuidade do cuidado contém regras geradas a partir de etiquetas morfológicas (a exemplo do trabalho de Bulegon), a fim de identificar as informações de continuidade do tratamento do paciente. Fez-se, então, necessária a execução do etiquetador (*part-of-speech tagger*) nos textos, para análise e busca manual de padrões de

continuidade do cuidado.

No Quadro 5 são exemplificados trechos que caracterizam a continuidade do cuidado pré-processados (etapa descrita a seguir) e etiquetados. Em negrito, os trechos que caracterizam continuidade do cuidado. As etiquetas são apresentadas ao lado das palavras, após o sinal de sublinhado ().

Os significados de todas as etiquetas que possam ser geradas a partir do etiquetador estão descritos na definição do *tagset* VISL Portuguese, disponível para acesso *online*, conforme listado nas referências. Exemplos de etiquetas utilizadas para encontrar trechos de continuidade do cuidado estão definidos no Quadro 6.

Quadro 5 - Trechos que caracterizam continuidade do cuidado etiquetados.

Trecho contendo continuidade	Trecho etiquetado
RECEBE ALTA BEM , DEVENDO SEGUIR ACOMPANHAMENTO COM MÉDICO ASSISTENTE.	recebe_V_PR_3S_IND_VFIN alta_N_F_S bem_ADV devendo_V_GER seguir_V_INF acompanhamento_N_M_S com_PRP médico_N_M_S assistente_ADJ_M_S ._-PNT_ABS
Recebe alta em bom estado geral, assintomática, com retorno ao ambulatório da cardiologia para acompanhamento.	recebe_V_PR_3S_IND_VFIN alta_N_F_S em_PRP bom_ADJ_M_S estado_N_M_S geral_ADJ_M_S assintomática_ADJ_F_S com_PRP retorno_N_M_S a_PRP o_DET_M_S ambulatório_N_M_S de_PRP a_DET_F_S cardiologia_N_F_S para_PRP acompanhamento_N_M_S ._-PNT_ABS
RECEBE ALTA EM BOM ESTADO GERAL, COM ORIENTAÇÃO DE MANTER ACOMPANHAMENTO COM SEU MÉDICO ASSISTENTE.	recebe_V_PR_3S_IND_VFIN alta_N_F_S em_PRP bom_ADJ_M_S estado_N_M_S geral_ADJ_M_S com_PRP orientação_N_F_S de_PRP manter_V_INF acompanhamento_N_M_S com_PRP seu_DET_M_S médico_N_M_S assistente_ADJ_M_S ._-PNT_ABS
PACIENTE ORIENTADA A PROCURAR MÉDICO ASSISTENTE PARA ACOMPANHAMENTO PÓS ALTA, E RETORNO À EMERGÊNCIA SE NECESSÁRIO.	paciente_N_M_S orientada_V_PCP_F_S procurar_V_INF médico_N_M_S assistente_ADJ_M_S para_PRP acompanhamento_N_M_S pós_ADJ_M_S alta_N_F_S e_KC retorno_V_PR_1S_IND_VFIN a_PRP a_DET_F_S emergência_N_F_S se_PERS_M/F_3S_ACC necessário_ADJ_M_S ._- PNT_ABS

Quadro 6 - Significados de etiquetas morfológicas utilizadas em padrões de continuidade do cuidado

Etiqueta	Significado
ADJ_M_S	Adjetivo Masculino Singular
ADV	Advérbio
DET_F_S	(Artigo Quantificador Determinante) Feminino Singular
DET_M_S	(Artigo Quantificador Determinante) Masculino Singular
KC	Conjunções Coordenativas
N_F_S	Substantivo Feminino Singular
N_M_S	Substantivo Masculino Singular
N_M_P	Substantivo Masculino Plural
NUM	Numeral
PRP	Preposição
V_INF	Verbo no Infinitivo
V_FUT_3S_IND_VFIN	Verbo no Futuro do Indicativo na 3ª pessoa do singular
V_GER	Verbo no Gerúndio
V_PR_3S_IND_VFIN	Verbo no Presente do Indicativo na 3ª pessoa do singular
V_PCP_M_S	Verbo no Particípio Masculino Singular

3.1.1 Análise dos sumários de alta

Antes de dar início ao processo de etiquetagem, os sumários de alta foram analisados e assinalados quanto à presença ou não de informação que caracterizasse continuidade do cuidado em meio ao texto. Essa tarefa foi realizada pelo Grupo de Pesquisa de Recuperação de Informação do PPGTS, que a partir da análise de alguns sumários, definiu o que caracterizava a descrição da continuidade do cuidado nos textos desta narrativa. A partir destas características uma aluna de iniciação científica do 10^a período do curso de medicina, participante do projeto, executou a indicação da presença ou ausência deste conteúdo em todos os 5617 sumários contidos no banco de dados. No 10^o período os alunos de medicina já elaboram textos que compõem os sumários e também tiveram a formação clínica necessária para conhecer as informações relevantes.

Além da informação sobre a continuidade do cuidado, também foram constatadas informações relevantes durante a leitura dos sumários de alta. Verificou-se que 69 sumários estavam em branco, ou seja, 1,23% do total já não teriam continuidade do cuidado e outras informações essenciais que devem ser descritas nos sumários de alta.

Foram marcados também os sumários que continham situação de óbito do paciente, totalizando, ao final, 205 textos. Sendo assim, esses 3,65% dos sumários não apresentavam informação de continuidade do cuidado.

Outra informação também relacionada à continuidade foi constatada em sumários que continham trechos semelhantes aos seguintes exemplos:

- Recebe alta em bom estado geral, com orientações;
- Orientado quanto a medicações e cuidados gerais;
- Orientada quanto a cuidados, medicações e retorno.

Nesses casos, percebe-se que as informações de continuidade de cuidado, podem ter sido passadas ao paciente de forma verbal, ou por meio de algum outro documento, porém, essas informações não são explícitas no texto do sumário, sendo assim, não é possível caracterizá-los como sumário de alta com orientações de continuidade do cuidado. No total, 259 sumários, ou 4,6%, continham esta característica.

Dos 5617 sumários de alta, 3127 deles continham alguma informação relacionada à continuidade de cuidado, ou seja, 55,67% do total.

3.1.2 Etiquetação

Para a etiquetação morfológica dos textos, o algoritmo da ferramenta CoGrOO foi treinado utilizando-se de um *corpus* médico anotado por Peters e seus colaboradores em 2010. Isso se deve ao fato do CoGrOO ter um *corpus* jornalístico embutido, no qual, não estão presentes palavras comuns ao vocabulário médico. Com a adoção do *corpus* médico evitam-se, portanto, erros de etiquetação nessas palavras.

Para testar a eficácia do etiquetador, foram selecionados 10 sumários de alta de maneira aleatória, e aplicado o algoritmo etiquetador. Percebeu-se uma diferença significativa no desempenho em sumários escritos com todas as letras maiúsculas, e nos que foram escritos corretamente, seguindo a norma formal de escrita, onde apenas as primeiras palavras das frases e nomes próprios têm letras maiúsculas.

Em média, os sumários escritos da maneira correta obtiveram **20% de erros** de etiquetação, enquanto os sumários escritos inteiramente em letras maiúsculas atingiram **87% de erros**. Essa alta taxa de erros pode ser explicada pelo uso do modelo de etiquetador estatístico, que compara a palavra lida no texto com as

palavras definidas no *corpus* de maneira *case-sensitive* (levando em conta diferença entre maiúsculas e minúsculas). Com isso, o algoritmo não encontra ocorrências de uma palavra inteiramente em maiúsculas no *corpus* e faz uma inferência da etiqueta de maneira errônea.

A partir dessa avaliação, concluiu-se que nos dois grupos de sumários (os escritos corretamente e aqueles com letras maiúsculas), a taxa de acertos não foi satisfatória o suficiente para utilização em um algoritmo que dependerá das etiquetas morfológicas para encontrar os padrões de continuidade do cuidado em meio ao texto. Assim sendo, a adição de uma fase de pré-processamento do texto, anteriormente à etiquetagem se fez necessária, para melhoria na taxa de acertos do etiquetador.

3.1.3 Pré-processamento

Dentre os problemas encontrados no texto, que resultaram em uma má etiquetagem, o que causou maior impacto negativo foram os textos escritos inteiramente em letras maiúsculas. Para correção, foi adicionado à fase de pré-processamento um método para normalizar as sentenças, colocando-as em letras minúsculas, evitando assim que o etiquetador não encontrasse referências às palavras do texto no *corpus* treinado.

Trechos de textos com a ocorrência de acrônimos, que são abreviaturas de palavras, com aparição constante em narrativas clínicas, foram responsáveis por erros de etiquetagem. Como solução, foi implementado um método de expansão de acrônimos, substituindo as abreviações pelo valor completo da palavra. Utilizou-se a mesma lista de acrônimos definida no método de Bulegon (2011).

Ao pré-processamento também foram adicionadas funcionalidades para retirar as *stop-words*, que são palavras de alta ocorrência no texto e que não tem valor semântico significativo, melhorando assim o desempenho do algoritmo, e também foram removidos alguns caracteres especiais e de pontuação que causavam problemas na etiquetagem do texto. A lista de *stop-words* ficou restrita aos artigos da língua portuguesa, já os caracteres removidos foram os definidos entre as aspas a seguir: “[!?!><=%;/#,@*]”.

Quadro 7 - Comparativo de sumário de alta etiquetado, com e sem a fase pré-processamento.

Sumário de alta original
#IAM INFERIOR 23/06 #HAS #TABAGISTA INTERNA PARA CATETERISMO DIAGNÓSTICO QUE REVELOU LESÃO EXCÊNTRICA 1/3 MÉDIO ACX ; SUBMETIDO A ACT + STENT DRIVER 2,5 X 12 MM NESTA LESÃO COM SUCESSO ANGIOGRÁFICO. RECEBE AS SEGUINTE RECOMENDAÇÕES NA ALTA: 1)USO DAS MEDICAÇÕES PRESCRITAS POR SEU MÉDICO 2)REVISÃO CLÍNICA COM SEU MÉDICO EM 7-10 DIAS 3)CLOPIDOGREL 75MG, VO, 1X/D POR 90 DIAS
Sumário de alta etiquetado sem fase de pré-processamento
#IAM_DET_M_S INFERIOR_PROP_M_S 23_NUM /06_NUM_M_P #HAS_N_M_P #TABAGISTA_PRP INTERNA_PROP_M/F_S PARA_-PNT_NSEP CATETERISMO_PROP_M_S DIAGNÓSTICO_-PNT_NSEP QUE_PROP_M/F_S REVELOU_V_PS_3S_IND_VFIN LESÃO_PRP EXCÊNTRICA_PROP_M/F_S 1_NUM /3_NUM_M_P MÉDIO_N_M_P ACX_ADJ_M_P ;_- PNT_REL SUBMETIDO_-PNT_ABS A_DET_F_S ACT_PROP_F_S +_\$+ STENT_PROP_M/F_S DRIVER_-PNT_ABS 2_NUM ,5_-PNT_ABS X_ADJ_F_S 12_NUM MM_PROP_F_S NESTA_PROP_F_S LESÃO_-PNT_ABS COM_PROP_M/F_S SUCESSO_-PNT_NSEP ANGIOGRÁFICO_PROP_M_S ._-PNT_ABS RECEBE_PROP_M/F_S AS_PRP SEGUINTE_PROP_M/F_S RECOMENDAÇÕES_-PNT_NSEP NA_PROP_M/F_S ALTA_V_PS_3S_IND_VFIN :_-PNT_REL 1_NUM_)_-PNT_REL USO_PRP DAS_PROP_M/F_S MEDICAÇÕES_PRP PRESCRITAS_PROP_M/F_S POR_-PNT_ABS SEU_PROP_M/F_S MÉDICO_-PNT_BIN 2_NUM_)_-PNT_REL REVISÃO_-PNT_ABS CLÍNICA_PROP_M/F_S COM_-PNT_ABS SEU_PROP_M/F_S MÉDICO_-PNT_ABS EM_PROP_M/F_S 7-10_-PNT_ABS DIAS_PROP_M/F_S 3_NUM_)_-PNT_REL CLOPIDOGREL_DET_F_S 75MG_PROP_F_S ,_- PNT_NSEP VO_PROP_F_S ,_-PNT_NSEP 1_NUM X_N_M_S /D_PRP POR_PROP_M/F_S 90_- PNT_ABS DIAS_PROP_M/F_S
Sumário de alta etiquetado com fase de pré-processamento
infarto_N_M_S agudo_ADJ_M_S de_PRP o_DET_M_S miocárdio_N_M_S inferior_ADJ_M_S 23_NUM de_PRP junho_N_M_S hipertensão_N_F_S arterial_ADJ_F_S sistêmica_ADJ_F_S tabagista_ADJ_F_S interna_N_F_S para_PRP cateterismo_N_M_S diagnóstico_ADJ_M_S que_SPEC_M_S revelou_V_PS_3S_IND_VFIN lesão_N_F_S excêntrica_ADJ_F_S 1_NUM 3_NUM médio_ADJ_M_S artéria_N_F_S coronariana_ADJ_F_S circunflexa_ADJ_F_S submetido_V_PCP act_ADV +_\$+ stent_N_M_S driver_N_M_S 2_NUM 5_NUM x_N_M_S 12_NUM mieloma_N_M_S múltiplo_ADJ_M_S em_PRP esta_DET_F_S lesão_N_F_S com_PRP sucesso_N_M_S angiográfico_ADJ_M_S ._-PNT_ABS recebe_V_PR_3S_IND_VFIN seguintes_ADJ_F_P recomendações_N_F_P em_PRP a_DET_F_S alta_N_F_S :_-PNT_REL 1_NUM_)uso_N_M_S de_PRP as_DET_F_P medicações_N_F_P prescritas_V_PCP_F_P por_PRP seu_DET_M_S médico_N_M_S 2_NUM_)_-PNT_REL revisão_N_F_S clínica_ADJ_F_S com_PRP seu_DET_M_S médico_N_M_S em_PRP 7_NUM 10_NUM dias_N_M_P 3_NUM_)_- PNT_REL clopidogrel_N_M_S 75mg_NUM_M/F_P via_PRP oral_ADJ_F_S 1x_-PNT_ABS d_N_M_S por_PRP 90_NUM_M_P dias_N_M_P

A presença de datas em meio ao texto também trouxe problemas de etiquetação. Foi, então, desenvolvido um método que por meio do uso de expressões regulares encontrasse e substituísse as datas no formato dd/mm/aaaa e mm/aaaa (Ex.: 27/02/1987 e 05/2003), para o formato escrito de forma corrente: dia de mês de ano (Ex.: 27 de fevereiro de 1987 e abril de 2003).

Algumas das etapas de pré-processamento descritas, foram apenas corrigidas durante as etapas posteriores de desenvolvimento do algoritmo, pois são correções decorrentes de problemas não encontrados inicialmente nos dez sumários selecionados.

Os dez sumários selecionados foram novamente submetidos à etiquetação, porém, com a fase de pré-processamento realizada anteriormente. Com isso, alcançou-se uma taxa média de **5,2% de erros**, indicando melhora significativa quando comparado com o texto sem o pré-processamento.

No Quadro 7 apresenta-se um exemplo de sumário de alta em sua forma original, etiquetado sem a fase de pré-processamento, e após, com o pré-processamento. Podem-se perceber situações de remoção de caracteres especiais (#), expansão de acrônimos (IAM expandido para Infarto Agudo do Miocárdio), etc.

3.2 DESENVOLVIMENTO DO ALGORITMO DE IDENTIFICAÇÃO DE CONTINUIDADE DO CUIDADO, *IRDISCHARGE*

O algoritmo utilizado para a identificação da continuidade do cuidado, nomeado de IRDischarge, foi definido através das fases e processos listados abaixo e abordados individualmente a seguir:

- 1) Pré-processamento
 - a. Expansão de acrônimos
 - b. Remoção de caracteres especiais
 - c. Normalização para letras minúsculas
 - d. Remoção de *stop-words*
 - e. Formatação das datas (escrita por extenso)
- 2) Processamento de Linguagem Natural
 - a. Detecção de sentenças
 - b. Tokenização
 - c. Detecção de nomes próprios

- d. Expansão de preposições
 - e. Etiquetagem morfológica
 - f. Corte e inversão das sentenças
- 3) Busca em meio ao texto etiquetado os padrões definidos nas regras
- a. Identificação das regras no texto.

Os processos executados na Fase 1 já foram abordados anteriormente, e são indispensáveis para o sucesso do algoritmo. Já os itens presentes na Fase 2 são as técnicas de PLN utilizadas no algoritmo, por meio da aplicação dos métodos já implementados pela ferramenta CoGrOO.

Conforme abordado na Revisão de Literatura, a Detecção de Sentenças segmenta o texto, separando-o em sentenças. Já a Tokenização separa todos os termos e pontuações encontrados em unidades únicas, para futuras avaliações morfológicas e sintáticas individuais.

Os dois processos seguintes, a Detecção de nomes próprios e a Expansão de Preposições, auxiliam em uma melhor etiquetagem das palavras. O primeiro identifica nomes próprios compostos em uma única unidade, por exemplo, o nome Pontifícia Universidade Católica do Paraná é definido como um substantivo composto. O principal motivo do uso da expansão de preposições é que o *corpus* anotado utilizado neste projeto também contém as preposições expandidas, fazendo com que os resultados de classificação morfológica sejam melhores com a utilização desta técnica. Além da possibilidade posterior de uso do Identificador de Sintagmas, que necessita das preposições expandidas para funcionar perfeitamente.

O etiquetador morfológico é executado, e por um processo estatístico de comparação com o *corpus* anotado utilizado, ele etiqueta as palavras morfológicamente com sua categoria principal (substantivo, verbo, adjetivo, etc.), e ainda faz uma inferência por meio regras pré-estabelecidas no CoGrOO para trazer as informações completas de flexões verbais (tempo, número, pessoa e modo), gênero (masculino, feminino) e número (singular, plural). A relação completa das categorias disponíveis na versão do CoGrOO utilizada está em VISL Portuguese, apresentado nas referências.

Na primeira etapa de geração de regras (descrita no próximo capítulo), verificou-se que geralmente a informação referente à continuidade do cuidado ao paciente encontrava-se nas últimas sentenças do texto.

Para reduzir a probabilidade do algoritmo encontrar um falso positivo, a ordem

de verificação das sentenças foi invertida, para que o algoritmo fizesse sua busca da última sentença para a primeira. Um levantamento estatístico foi feito (vide Tabela 1), e verificou-se que dificilmente a informação de continuidade do cuidado estava presente fora das 4 últimas sentenças. Com essa constatação, o algoritmo, então, removeria todas as sentenças fora do intervalo das 4 últimas sentenças, evitando muitas prováveis ocorrências de falsos positivos.

Tabela 1 – Frequência de ocorrência de continuidade de cuidado por sentenças após a inversão na ordem das mesmas.

Sentença	Frequência de ocorrência de continuidade do cuidado
1 ^a	64%
2 ^a	24%
3 ^a	9%
4 ^a	2%
9 ^a	1%

Como último processo do algoritmo, pertencente à Fase 3, é feita a tentativa de encontrar em meio ao texto etiquetado a ocorrência de um padrão definido em alguma das regras do sistema e, assim que encontrada alguma ou nenhuma ocorrência, o algoritmo encerra a análise do sumário de alta e retorna o resultado. Importante ressaltar que o algoritmo tem uma ordem específica de verificação das regras, que quando inseridas no banco de regras, tem a informação da ordem de verificação da mesma.

3.3 METODOLOGIA DE DESENVOLVIMENTO DE REGRAS

O processo de definição das regras foi executado de forma manual, a partir da análise dos sumários etiquetados. Inicialmente foi um trabalho conjunto visando dois projetos distintos, este e o trabalho de Andréia Cristina de Souza (SOUZA, 2012). Porém, uma diferenciação na definição de quais informações seriam consideradas como continuidade do cuidado ao paciente, fez com que fosse necessária uma nova etapa de geração de regras, diferente da feita no projeto de SOUZA, 2012.

A diferença básica com o trabalho de SOUZA é que além de informações como orientações de continuidade de tratamento (exceto medicamentoso) e

encaminhamento para outro profissional ou especialidade, também foram considerados como continuidade do cuidado os dados de retorno pós-alta, que elevaram significativamente a quantidade de informações a serem localizadas.

Dos 5617 sumários de alta, 3127 deles tiveram ocorrências de continuidade do cuidado, entre eles, 1441 são caracterizados por informações de retorno pós-alta, ou seja, um aumento 46,1% no número ocorrências de continuidade do cuidado, se comparado com o trabalho de SOUZA, 2012. No Quadro 8, são apresentados exemplos de informações de retorno pós-alta em trechos de sumários de alta.

Visando a utilização do menor número possível dos 5617 sumários de alta disponíveis e, conseqüentemente, um desenvolvimento de regras mais rápido, foi necessária a especificação de uma metodologia para definição do banco de regras.

A metodologia de desenvolvimento das regras pode ser resumida em duas etapas principais, enumeradas a seguir:

- 1) Etapa de desenvolvimento inicial de regras;
- 2) Etapa de desenvolvimento iterativo de regras.

Quadro 8 – Exemplos de informações de retorno pós-alta

Trechos de sumários de alta	Trechos que caracterizam retorno pós-alta
Tem alta com plano de retorno ambulatorial breve no ambulatório de ICC. Avaliar possibilidade de anticoagulação ambulatorialmente	plano de retorno ambulatorial breve no ambulatório de ICC.
RECEBE ALTA EM BOM ESTADO GERAL, COM ORIENTAÇÃO DE MANTER MEDICAÇÕES EM USO. RETORNO AO CONSULTÓRIO DO SEU MÉDICO ASSISTENTE EM 7-10 DIAS.	RETORNO AO CONSULTÓRIO DO SEU MÉDICO ASSISTENTE EM 7-10 DIAS.
Recebe alta em boas condições clínicas com retorno para o ambulatório de cardiologia do hospital em 04/10/02.	retorno para o ambulatório de cardiologia do hospital em 04/10/02.

3.3.1 Etapa de desenvolvimento inicial de regras

Como ponto de partida para a avaliação de desempenho do algoritmo, definiu-se um banco inicial de regras. Para tal tarefa, foram selecionados aleatoriamente 100 textos dentre os sumários de alta marcados com a presença de continuidade do cuidado, e executou-se o pré-processamento e etiquetagem dos

mesmos.

A partir da leitura dos 100 textos etiquetados, foram encontrados manualmente em meio ao texto, padrões que indicam a continuidade de cuidado, totalizando, ao final, 40 regras. Este conjunto de regras é utilizado como base para iniciar a etapa de desenvolvimento iterativo de regras, denominado conjunto inicial. No Quadro 9, são mostrados exemplos das 4 regras geradas que somaram mais ocorrências.

Quadro 9 – Exemplos de regras geradas na etapa de desenvolvimento inicial de regras

Trechos de sumários de alta etiquetados	Regras extraídas
manter_V_INF acompanhamento_N_M_S ambulatorial_ADJ_M_S e_KC retornar_V_INF a_PRP a_DET_F_S emergência_N_F_S em_PRP caso_N_M_S de_PRP piora_N_F_S	<V_INF><N_M_S><ADJ_M_S>
tem_V_PR_3S_IND_VFIN retorno_N_M_S a_PRP o_DET_M_S ambulatório_N_M_S de_PRP medicina_N_F_S interna_ADJ_F_S	<N_M_S><PRP><DET_M_S> <N_M_S><PRP>
manter_V_INF acompanhamento_N_M_S em_PRP o_DET_M_S ambulatório_N_M_S de_PRP cardiologia_ADJ_M_S	<V_INF><N_M_S><PRP>
com_PRP retorno_N_M_S a_PRP o_DET_M_S seu_DET_M_S médico_N_M_S assistente_ADJ_M_S para_PRP acompanhamento_N_M_S clínico_ADJ_M_S	<N_M_S><PRP><DET_M_S> <DET_M_S>

As 40 regras geradas foram inseridas no banco de regras do algoritmo IRDischarge. A ordenação das mesmas foi definida por dois critérios, na ordem descendente: frequência de aparecimento da regra e quantidade de termos da regra. Quanto maior o valor desses dois fatores, mais acima a regra ficaria no banco de regras, ou seja, antes ela seria verificada na execução do algoritmo.

Tais parâmetros de ordenação foram definidos pelo fato de uma regra que teve muitas ocorrências positivas tem maior probabilidade de trazer resultados positivos em iterações executadas futuramente, e uma regra com maior quantidade de termos tem menor probabilidade de trazer falsos positivos por ser mais específica e de padrão mais difícil de ser repetido durante o texto.

Na Figura 2, ao lado esquerdo, em verde, são apresentados os processos executados nesta etapa inicial de desenvolvimento de regras.

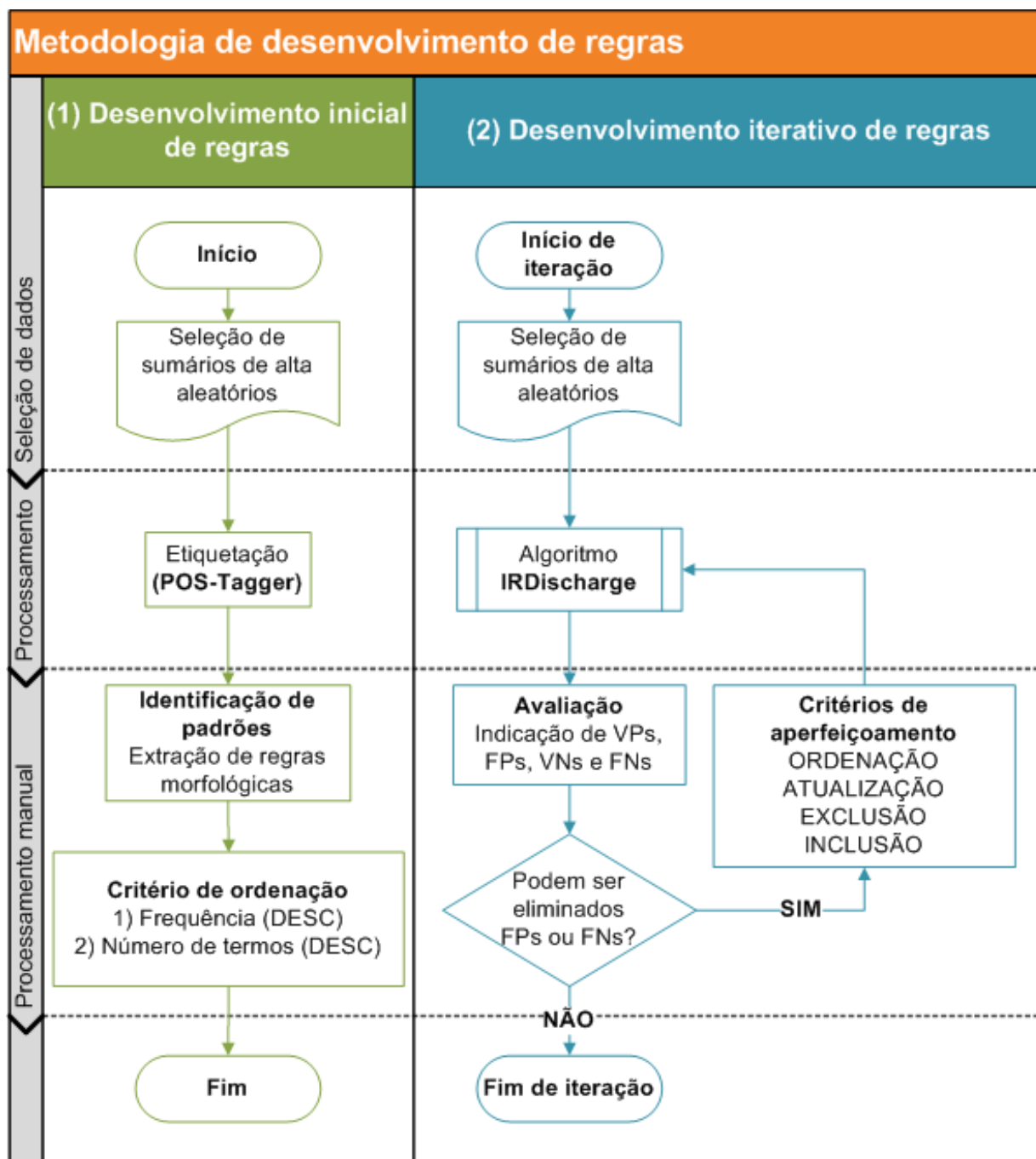


Figura 2 – Etapas da metodologia de desenvolvimento de regras

3.3.2 Etapa de desenvolvimento iterativo de regras

Com as 40 primeiras regras definidas, deu-se início a etapa iterativa de desenvolvimento das regras, onde o objetivo é, a cada nova iteração, efetuar melhorias no banco de regras em busca do conjunto de regras ideal para o algoritmo IRDischarge.

É importante ressaltar que todos os sumários selecionados a partir da etapa

inicial não são reaproveitados durante as iterações posteriores, e são retirados do banco de dados ao final de cada iteração, portanto, os 100 sumários de alta selecionados na etapa de desenvolvimento inicial de regras já foram descartados.

Na Figura 2, à direita, na cor azul, são apresentados os processos executados na etapa iterativa de desenvolvimento de regras, detalhados a seguir.

Inicialmente, são selecionados aleatoriamente 100 sumários de alta, porém, ao contrário da etapa inicial de desenvolvimento de regras, não são selecionados apenas sumários dentre os que contêm continuidade do cuidado, e sim entre o conjunto completo de sumários de alta.

Após a seleção, os textos são passados ao algoritmo IRDischarge, que por meio do uso das regras contidas no banco de regras, busca ocorrências de continuidade do cuidado nos sumários de alta, e efetua uma saída com a resposta do algoritmo no documento, pontuando as informações encontradas, ou informando a ausência de informações relevantes.

Através dos documentos marcados com as respostas do algoritmo IRDischarge, foi possível avaliar os resultados do processamento dos textos, fazendo um levantamento de Verdadeiros Positivos (VPs), Falsos Positivos (FPs), Verdadeiros Negativos (VNs) e Falsos Negativos (FNs). O Quadro 10 apresenta trechos de sumários processados, com as respostas do IRDischarge, bem como, a definição do resultado, feita de maneira manual após o processamento, para avaliação geral da iteração.

Após análise individual dos resultados dos 100 sumários, verifica-se a possibilidade de correção dos erros (remover ocorrências de FPs e FNs). As ações possíveis para correção de um erro são:

- 1) Alterar ordem de execução da regra (**ORDENAÇÃO**);
- 2) Atualizar regra (reduzir ou aumentar quantidade de termos) (**ATUALIZAÇÃO**);
- 3) Retirar regra do banco de regras (**EXCLUSÃO**);
- 4) Inserir nova regra no banco de regras (**INCLUSÃO**).

Para ordenação das regras, foram estabelecidos critérios apontados na Figura 3.

Como primeiro **fator de ordenação** efetua-se uma avaliação acerca do sucesso geral da regra, ou seja, se ela obtiver uma melhor relação entre VPs e FPs em uma posição, se comparada a ela em outra posição, sempre será mantida na

posição que obteve mais sucesso.

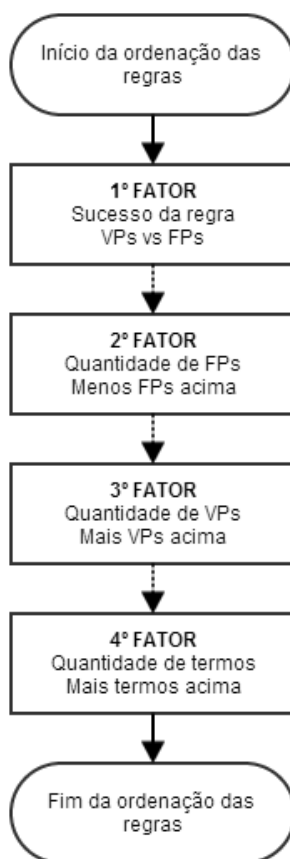


Figura 3 – Critérios para ordenação das regras

A quantidade de FPs de uma regra é o segundo fator de ordenação, regras que têm menor probabilidade de FPs são mantidas nas primeiras posições, visto que quando o algoritmo IRDischarge for executado sempre dará prioridade às regras que tem poucas chances de causar resultados negativos à iteração atual.

O número de VPs de uma regra é o terceiro fator de ordenação, assim, teremos as regras que não causam erros ao algoritmo a frente (menos FPs), seguidas das regras que mais trazem resultados positivos (mais VPs).

Finalizando com o último fator de ordenação, que é a quantidade de termos da regra, quanto mais termos, mais acima ela é posicionada. Este critério funciona com um “critério de desempate” para o posicionamento de regras que contém números iguais de VPs e FPs. Esse critério foi estabelecido visto que as regras com mais termos tornam-se mais específicas e mais difíceis de serem encontradas em meio ao texto, portanto, tem possibilidade estatística reduzida de retornarem um FP.

Quadro 10 – Trechos de sumários processados pelo IRDischarge e resultado do processamento

Trechos de sumários de alta processados	Resposta do IRDischarge	Definição do resultado
RECOMENDAÇÕES NA ALTA: 1) TOMAR MEDICAÇÕES CONFORME PRESCRIÇÃO 2) CONSULTAR COM SEU MÉDICO EM BREVE 3) RETORNAR A EMERGÊNCIA SE NECESSÁRIO	TRECHO ENCONTRADO: consultar com seu médico REGRA: <V_INF><PRP><DET_M_S><N_M_S>	VP
Paciente submetido a angioplastia coronariana com implante de stent em segmento proximal de artéria coronariana direita. Procedimento transcorreu sem intercorrências.	NENHUM TRECHO ENCONTRADO.	VN
Paciente sendo transferido para o Hospital Universitário para ser submetido a angioplastia com implante de stent. Aqui no Hospital não foi possível realizar o procedimento por problemas técnicos, falta de material específico para este procedimento.	TRECHO ENCONTRADO: realizar procedimento por REGRA: <V_INF><N_M_S><PRP>	FP
RECEBE ALTA EM BOAS CONDIÇÕES CLÍNICAS, COM INFECÇÃO TRATADA, SEM CONGESTÃO E COM PLANO DE SEGUIMENTO AMBULATORIAL NESTA INSTITUIÇÃO.	NENHUM TRECHO ENCONTRADO.	FN

A **atualização de termos** de uma regra é feita quando é vista a necessidade de especializar a regra (adicionar termos) ou generalizar a mesma (retirar termos). O Quadro 11 apresenta exemplos de especialização e generalização de regras.

Geralmente, a especialização é necessária quando a regra ocasiona muitos FPs, mesmo em uma posição mais abaixo na ordenação. Nessa situação, os trechos de texto onde há ocorrência da regra são analisados e verifica-se se é possível adicionar mais termos a ela para redução dos FPs.

Já a generalização é feita para tentar aumentar a quantidade de VPs da regra. Analisando os textos processados na iteração, pode-se perceber em alguns casos que se a regra fosse reduzida, poderia obter outros padrões de continuidade

do cuidado não cobertos em sua forma atual, aumentando, assim, a quantidade de ocorrências corretas para o algoritmo.

Em situações em que a regra já foi reordenada, atualizada (generalização ou especialização) e continua produzindo mais resultados negativos que positivos, opta-se por retirá-la do banco de regras.

Todos FNs gerados ao final do processamento são analisados, e os padrões não encontrados devem encaixar-se em alguma regra já existente (generalizando ou especializando), ou, em não havendo essa possibilidade, uma nova regra é definida para o padrão de continuidade de cuidado.

Depois de concluídas as quatro etapas (ordenação, atualização, remoção e inserção de regras), mantém-se o mesmo conjunto de sumários de alta selecionados no início na iteração, e um novo processamento do algoritmo IRDischarge é executado para nova avaliação.

Quadro 11 – Exemplos de Especialização e Generalização de regras

Regra	Atualização da regra
<N_M_S><PRP><DET_M_S><N_M_S> Em determinada iteração trouxe 24 FPs Exemplo de trecho encontrado: Reencaminhado a pedido de o paciente.	ESPECIALIZAÇÃO <N_M_S><PRP><DET_M_S><N_M_S> <PRP> Após especialização trouxe 12 FPs
<V_PR_3S_IND_VFIN><N_F_S><PRP><N_M_S> <PRP> Em determinada iteração trouxe 2 VPs Exemplo de trecho encontrado: recebe alta para seguimento com seu médico	GENERALIZAÇÃO <V_PR_3S_IND_VFIN><N_F_S><PRP><N_M_S> Após generalização trouxe 4 VPs Exemplo de padrão adicionado: recebe alta para seguimento ambulatorial

No momento em que não há mais possibilidade de melhorias no banco de regras, as métricas de Recuperação de Informação (Precisão, Revocação, Especificidade e F-Measure) são calculadas e guardadas, e uma nova iteração é iniciada, com novos sumários de alta e mais avaliações de resultados e melhorias no banco de regras.

A geração de regras só chega ao seu fim, quando os valores das métricas calculadas ao fim de cada iteração se mantêm estáveis, mesmo sem terem sido feitas melhorias após o primeiro processamento dos textos da iteração. Portanto, se durante 3 iterações consecutivas não houve intervenções no banco de regras, e os valores das métricas se mantiveram estáveis, é dado o fim da geração de regras,

conforme Figura 4.

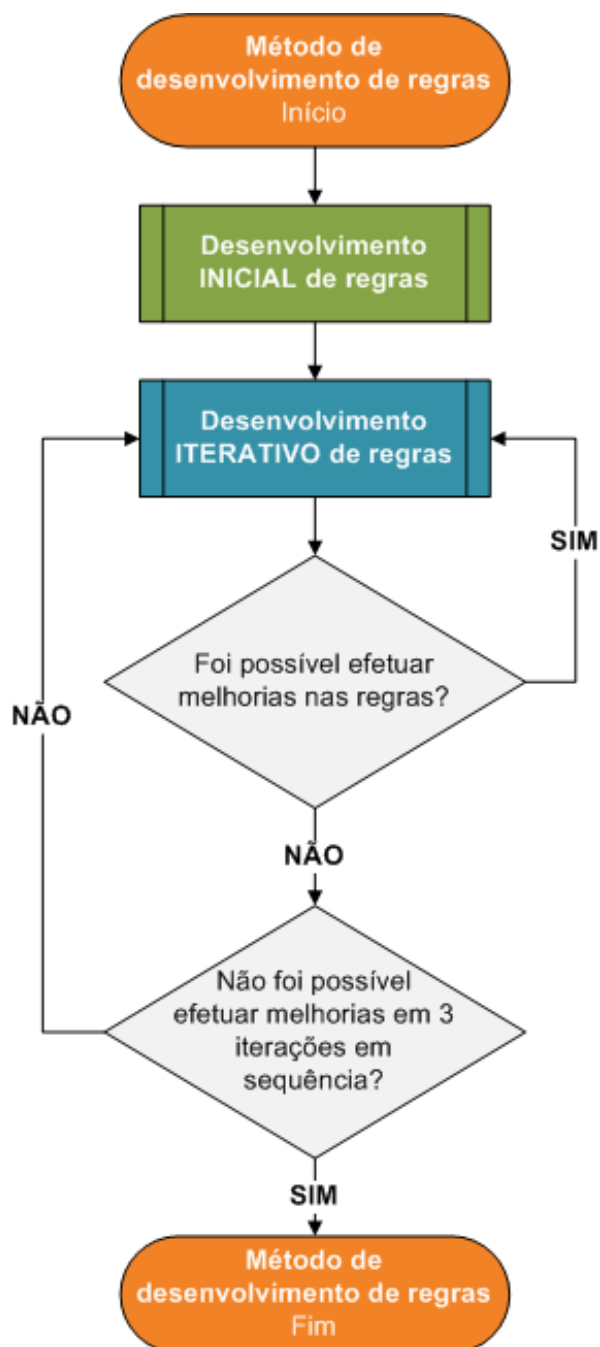


Figura 4 – Visão ampla dos processos de Geração das regras

Além da etapa de desenvolvimento inicial de regras, houve mais 7 iterações da etapa de desenvolvimento iterativo de regras, para refinamento do banco de regras. A Tabela 2 demonstra as iterações e a quantidade de regras definidas ao final do último processamento de cada iteração. As últimas 3 iterações tiveram apenas um processamento e nenhuma intervenção no banco de regras. A iteração marcada como “Conjunto inicial”, caracteriza os sumários selecionados na etapa inicial de

desenvolvimento de regras.

Tabela 2 – Iterações executadas para geração de regras

Iteração	Sumários processados	Regras definidas
Conjunto inicial	100	40
1	100	17
2	100	24
3	100	27
4	100	30
5	100	30
6	100	30
7	100	30

A quantidade de dados (sumários de alta) necessários para **treinamento** do algoritmo, seguindo a metodologia de geração de regras proposta, é definida de maneira dinâmica e não fixa no início do processo como normalmente ocorre em projetos de pesquisa de RI, podendo variar bastante de acordo com o domínio de utilização. Neste caso, totalizaram-se 500 sumários de alta para treinamento. Já a quantidade de dados para **testes** se resume às últimas 3 iterações onde não existem alterações no banco de regras, totalizando 300 sumários de alta para testes. Na Tabela 2, em negrito, estão as iterações utilizadas para teste do algoritmo.

3.4 AVALIAÇÃO DE DESEMPENHO DO ALGORITMO

Para a avaliação do algoritmo de identificação da continuidade do cuidado, foram efetuados cálculos específicos utilizados em sistemas de RI.

Ao final de cada iteração de desenvolvimento de regras, são levantados os seguintes valores para realizar a avaliação:

- VP (Verdadeiro positivo): o sumário contém informações de continuidade do cuidado e o algoritmo identificou a informação;
- FP (Falso positivo): o sumário não contém informações de continuidade do cuidado, porém, o algoritmo identificou de forma inadequada (É também considerado FP, quando existe informação de continuidade, porém o algoritmo identifica outro trecho de texto dentro do sumário);
- VN (Verdadeiro negativo): o sumário não contém informações de

continuidade do cuidado e o algoritmo não identificou nenhuma informação;

- FN (Falso negativo): o sumário contém informações de continuidade do cuidado, porém, o algoritmo não identificou a informação.

Dado que esta pesquisa tem como foco explicitar ao profissional da saúde os sumários de alta que não contém informações vitais ao tratamento, como a continuidade do cuidado, além do cálculo da **Precisão**, **Revocação** e **F-Measure**, é interessante também calcular a **Especificidade** do algoritmo, que é a habilidade do sistema identificar resultados negativos, ou seja, a ausência da informação de continuidade do cuidado.

Como o fim da geração de regras é dado quando os valores mantêm-se estáveis durante as três últimas iterações sem intervenções nas regras, a avaliação final de desempenho do algoritmo será uma média aritmética de cada um dos valores nas três iterações finais de processamento.

3.5 DESENVOLVIMENTO DE FERRAMENTA PARA APOIO À ESCRITA DE SUMÁRIOS DE ALTA

Através do algoritmo (IRDischarge) e banco de regras desenvolvidos, uma ferramenta para auxílio a escrita do sumário de alta foi construída para evidenciar ao profissional médico acerca da completude do texto que está sendo escrito.

O objetivo é que a ferramenta seja modular e acoplável, para futuramente aceitar a inserção dos algoritmos que identifiquem os outros elementos-chave (definidos por Souza, 2012) dentro do sumário de alta, não apenas a continuidade do cuidado.

Optou-se pela construção de uma ferramenta *web-based*, visto que é acessível facilmente de qualquer dispositivo conectado à *Internet* (*PCs*, *notebooks*, *tablets* e celulares). A linguagem escolhida para o desenvolvimento *server-side* foi o JAVA, já que o IRDischarge e CoGrOO utilizam-se da mesma linguagem de programação. Para os scripts *client-side*, utilizou-se a biblioteca JavaScript, jQuery.

Para evitar erros ortográficos por parte do usuário durante a digitação do sumário de alta, um corretor ortográfico foi acoplado ao navegador utilizado. Com isso pretende-se evitar alguns erros de etiquetação provindos de erros ortográficos.

O funcionamento básico da ferramenta é simples, o usuário tem acesso a um campo de entrada de dados para digitar o sumário de alta. Ao lado direito, há um

quadro com todos os elementos-chave do sumário de alta. Durante a digitação do texto o sistema identifica por meio de cores (vermelho para não preenchido e verde para preenchido) o *status* do elemento-chave atual dessa pesquisa, que é a continuidade do cuidado.

4 RESULTADOS

São resultados deste trabalho o IRDischarge, algoritmo para identificação da continuidade do cuidado em sumários de alta, incluindo as regras geradas para o seu funcionamento eficiente e a metodologia proposta para a geração das mesmas; a avaliação de desempenho do algoritmo e a ferramenta de apoio à escrita de sumários de alta.

4.1 ALGORITMO DE IDENTIFICAÇÃO DE CONTINUIDADE DO CUIDADO, IRDISCHARGE

O IRDischarge é composto por três passos principais, que são representados nos retângulos maiores do fluxograma apresentado na Figura 5. Esses passos são aplicados em cada sumário de alta em que se deseja identificar a existência de continuidade do cuidado ao paciente.

O primeiro passo é o pré-processamento do texto do sumário de alta, incluindo cinco etapas a serem realizadas na seguinte sequência: expansão de acrônimos, remoção de caracteres especiais, normalização para letras minúsculas, remoção de *stop-words* e formatação de datas.

O segundo passo, Processamento de Linguagem Natural, corresponde à aplicação das seguintes técnicas nos textos: detecção de sentenças, tokenização, expansão de preposições, detecção de nomes próprios e etiquetagem morfológica, disponíveis no CoGrOO, e que obrigatoriamente devem ser aplicadas nessa ordem.

O processamento dos textos dos sumários de alta é o último passo, no qual as sentenças do texto tem sua ordem invertida, ou seja, a última sentença passa a ser a primeira, e são mantidas apenas as quatro primeiras sentenças. A seguir, o banco de regras é consultado e, seguindo a ordem pré-estabelecida, uma a uma, são buscadas ocorrências da regra em meio ao texto. Quando é encontrada uma ocorrência, o algoritmo retorna como resultado o trecho encontrado e a regra utilizada para encontrá-lo. Caso todas as regras tenham sido avaliadas, e nenhuma ocorrência tenha sido encontrada, como resultado o algoritmo informa a ausência da continuidade do cuidado no sumário de alta em questão.

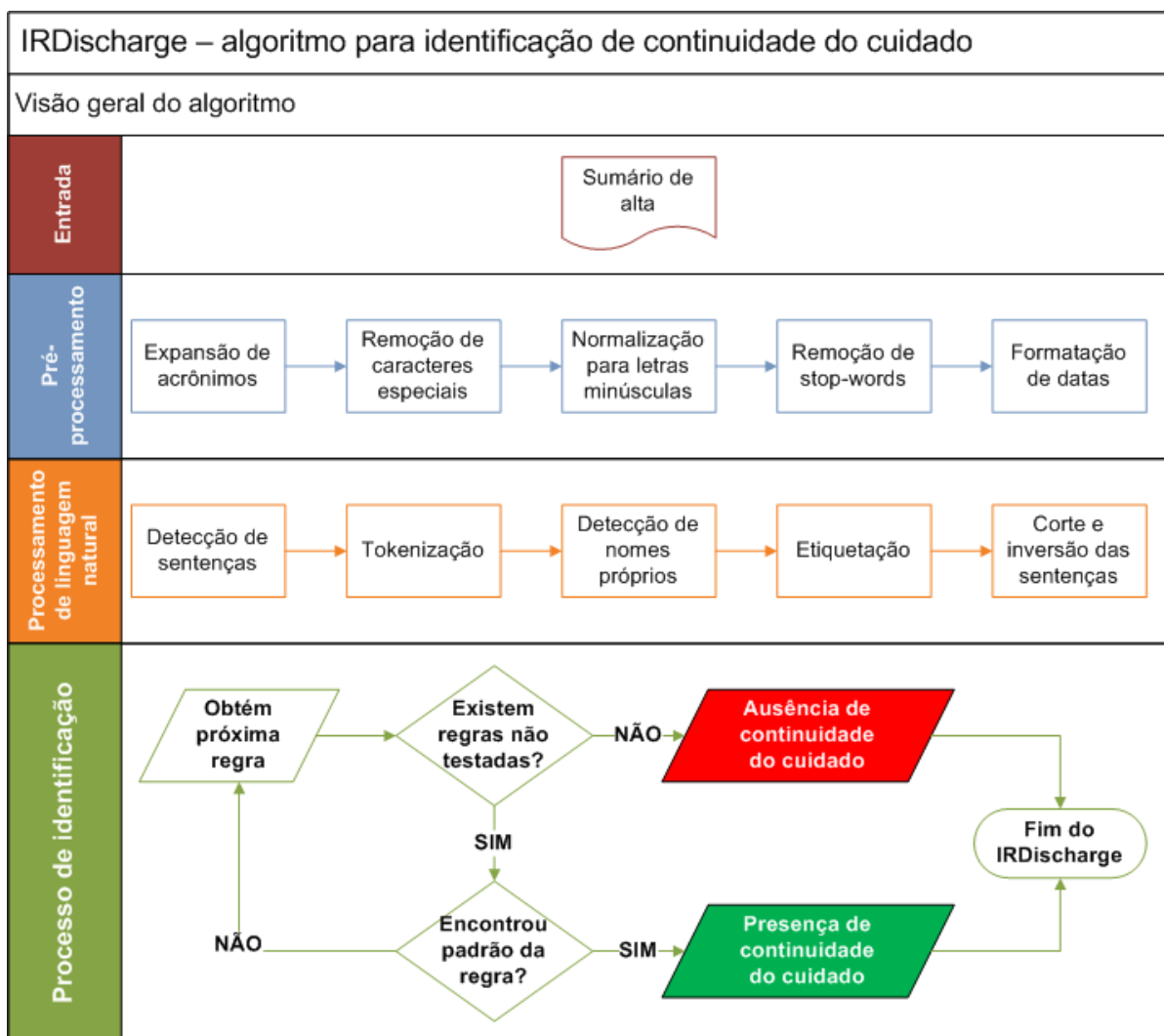


Figura 5 - Fluxo de processos para identificação da continuidade do cuidado (IRDischarge)

4.2 BANCO DE REGRAS DO ALGORITMO

Através do uso da Metodologia de Desenvolvimento de regras, foram definidas trinta regras finais, baseadas em etiquetas morfológicas, que caracterizam padrões que identificam a continuidade do cuidado, as quais são apresentadas no Quadro 12. Para cada regra citam-se exemplos de trechos dos sumários que caracterizam sua ocorrência. Vale ressaltar que os textos dos exemplos estão na forma original, sem a expansão de preposições, que é executada pelo algoritmo IRDischarge. Deste modo, no primeiro trecho da Regra 2 por exemplo, a palavra “ao”, após a expansão, é separada em duas palavras “a” (PRP) e “o” (DET_M_S). Assim, a regra que identifica este trecho é composta por cinco termos e a frase original contém quatro termos.

Quadro 12 - Regras finais para identificação da continuidade do cuidado

Ordem	Regra
1	<N_F_S><ADJ_F_S><PRP><DET_M_S><N_M_S><PRP><NUM> <u>Revisão clínica com seu médico em 7 dias.</u> <u>Revisão clínica com seu médico em 7-10 dias.</u>
2	<V_INF><PRP><DET_M_S><N_M_S> <u>Retornar ao ambulatório</u> de insuficiência cardíaca. <u>Retornar ao médico</u> assistente. <u>Consultar com seu médico</u> assistente em 7-10 dias. <u>Retornar ao hospital</u> de sua cidade.
3	<V_GER><N_M_S><PRP><N_M_S><ADJ_M_S> <u>Seguindo acompanhamento com médico</u> assistente.
4	<N_M_S><ADJ_M_S><PRP><DET_M_S><DET_M_S> <u>Retorno breve ao seu</u> médico para revisão. <u>Retorno imediato ao seu</u> médico
5	<V_INF><DET_M_S><N_M_S><PRP><NUM><N_M_P> <u>Consultar seu médico em 10 dias.</u>
6	<V_INF><PRP><NUM> <u>Retornar em 2</u> semanas para realizar ACTP+stent.
7	<N_M_S><PRP><DET_M_S><N_M_S><ADJ_M_S> <u>Retorno ao médico</u> assistente. <u>Retorno com seu médico</u> assistente. <u>Retorno ao consultório</u> médico. <u>Retorno para seu médico</u> assistente. <u>Manter medicações e acompanhamento com seu médico</u> assistente.
8	<V_INF><N_M_S><ADJ_M_S> <u>Seguir acompanhamento ambulatorial</u> com seu cardiologista. Plano de <u>manter seguimento ambulatorial</u>. Com orientação de <u>manter acompanhamento ambulatorial</u>. <u>Procurar clínico geral</u> ou cardiologista para acompanhamento. <u>Retorno no ambulatório da ICC para seguir acompanhamento clínico.</u> Optou-se, então, por <u>manter tratamento clínico</u>. <u>Consultar médico</u> assistente em uma semana. <u>Manter tratamento conservador</u> para fraturas costais. <u>Manter acompanhamento regular</u> com cardiologista em sua cidade. Tem alta para <u>manter tratamento prévio</u>.
9	<N_M_S><PRP><DET_M_S><DET_M_S> <u>Retorno ao seu</u> médico para revisão e acompanhamento. <u>Retorno ao seu</u> médico assistente. <u>Retorno ao seu</u> médico para seguimento.

10	<N_M_S><PRP><DET_M_S><N_M_S><PRP>
	<u>Retorno ao Ambulatório de ICC para acompanhamento.</u>
	<u>Retorno ao Ambulatório do hospital para seguir tratamento.</u>
	<u>Retorno ao médico de origem.</u>
11	<N_M_S><PRP><NUM><N_M_P><PRP>
	<u>Retorno em 7 dias com seu médico assistente.</u>
	Com plano de <u>retorno em 4 dias no ambulatório de arritmias.</u>
	<u>Retorno em 10 dias para revisão.</u>
12	<V_INF><N_M_S><PRP>
	<u>Seguir acompanhamento com seu cardiologista.</u>
	<u>Manter acompanhamento com seu médico.</u>
	<u>Procurar médico em sua cidade para controle de hipertensão.</u>
13	<N_M_S><ADJ_M_S><PRP><NUM><N_M_P>
	Revisão com seu <u>médico assistente em 7 dias.</u>
14	<PRP><DET_M_S><N_M_S><ADJ_M_S><PRP><N_M_S>
	<u>Ao médico assistente para seguimento.</u>
15	<N_M_S><PRP><N_M_S><ADJ_M_S><PRP><N_F_S>
	<u>Plano de tratamento farmacológico para cardiopatia.</u>
	<u>Plano de seguimento ambulatorial com cardiologia.</u>
	<u>Retorno com médico assistente para reavaliação.</u>
16	<V_PR_3S_IND_VFIN><N_F_S><N_M_S><ADJ_M_S>
	<u>Segue investigação nível ambulatorial.</u>
17	<PRP><N_M_S><V_PCP_M_S><PRP><N_F_S>
	<u>Com retorno agendado com neurologia deste hospital.</u>
	<u>Com retorno marcado para consulta ambulatorial.</u>
18	<V_FUT_3S_IND_VFIN><V_PCP_M_S><ADV>
	<u>Será acompanhado ambulatorialmente pela nefro.</u>
19	<KC><PRP><DET_M_S><DET_M_S><N_M_S><PRP><N_M_S>
	Alta e <u>ao seu médico para seguimento.</u>
20	<V_PR_3S_IND_VFIN><N_F_S><PRP><N_M_S>
	<u>Recebe alta para seguimento com seu médico.</u>
	<u>Recebe alta com encaminhamento para o ambulatório da cardiologia.</u>
	<u>Recebe alta com plano de prosseguir ambulatorialmente a investigação.</u>
21	<V_PR_3S_IND_VFIN><N_M_S><ADJ_M_S>
	Paciente <u>segue tratamento clínico</u> com seu médico assistente.
	<u>Tem retorno ambulatorial</u> marcado para o dia 27/03/2007.
	<u>Tem retorno breve ambulatorial.</u>
22	<V_GER><N_M_S><PRP><DET_M_S><N_M_S>

	<u>Seguindo acompanhamento com seu médico.</u>
23	<N_F_S><PRP><DET_M_S><N_M_S><PRP><N_F_S> <u>Reavaliação no hospital de origem.</u> <u>Consulta pós alta no ambulatório de cardiologia geral.</u> <u>Revisão no ambulatório de Cardiologia.</u>
24	<PRP><N_M_S><PRP><V_INF> <u>Com recomendação de seguir acompanhamento com seu médico.</u> <u>Com plano de seguir acompanhamento no ambulatório de Cardiologia.</u> <u>Com plano de retornar em janeiro para ACTP.</u>
25	<V_INF><N_F_S><PRP><N_M_S><PRP><DET_F_S><N_F_S> <u>Marcar consulta de retorno na cardiologia.</u>
26	<N_M_S><ADJ_M_S><PRP><NUM><NUM><N_M_P> <u>Retorno ao médico assistente em 7-10 dias.</u>
27	<V_GER><V_INF><N_F_S> <u>Devendo realizar marcação cirúrgica ambulatorialmente.</u> <u>Devendo manter terapia ambulatorialmente.</u>
28	<V_PR_3S_IND_VFIN><PRP><N_M_S><ADJ_M_S><PRP><DET_M_S><N_M_S> <u>Permanece em acompanhamento ambulatorial neste hospital</u>
29	<KC><N_M_S><ADJ_M_S><PRP><DET_M_S><N_M_S> <u>Decidido por tto clínico e seguimento ambulatorial com seu médico.</u> <u>Orientações quanto ao seguimento e acompanhamento ambulatorial com seu médico.</u> <u>Uso das medicações habituais e acompanhamento regular com seu médico.</u> <u>Reavaliação em breve e acompanhamento regular com seu cardiologista.</u> <u>Optando-se por manejo conservador e seguimento ambulatorial com seu médico.</u>
30	<V_PCP_M_S><V_INF><N_F_S> <u>Orientado a procurar cardiologista e nefrologista.</u>

4.3 AVALIAÇÃO DE DESEMPENHO DO ALGORITMO

A avaliação de desempenho do algoritmo de identificação da continuidade do cuidado foi definida utilizando-se das metodologias próprias para mensuração de sistemas de RI. Na Tabela 3, descrevem-se os valores obtidos ao final do primeiro processamento, em cada uma das iterações de geração de regras.

É importante lembrar que se considera também FP quando o sumário de alta contém continuidade do cuidado, porém, alguma regra encontra um trecho que não representa essa informação dentro deste sumário.

Tabela 3 - Resultado do processamento dos sumários quanto às informações encontradas

Iteração	Ausência de Continuidade	Presença de Continuidade	VP	FP	VN	FN
1	46	54	35	14	38	13
2	41	59	32	14	38	16
3	30	70	51	5	27	17
4	48	52	31	6	46	17
5	40	60	51	4	36	7
6	44	56	47	3	41	9
7	39	61	52	4	37	7

A partir dos dados da Tabela 3, foram calculados os valores correspondentes a Precisão, Revocação, Especificidade e F-Measure do sistema, apresentados na Tabela 4.

Tabela 4 – Avaliação do processamento dos sumários de alta por iteração

Iteração	Precisão	Revocação	Especificidade	F-Measure
1	71%	73%	73%	72%
2	70%	67%	73%	68%
3	91%	75%	84%	82%
4	84%	65%	88%	73%
5	93%	88%	90%	90%
6	94%	84%	93%	89%
7	93%	88%	90%	90%

Na Figura 6, ilustra-se o gráfico de desempenho do algoritmo ao final do primeiro processamento de cada iteração de desenvolvimento de regras.

As curvas indicam uma leve variação dos resultados nas três últimas iterações, nos quais, não houve possibilidade de intervenção no banco de regras para melhoria dos resultados. Devido a isso, não foram realizadas mais iterações.

Como resultado final (apresentado na Tabela 5), foi calculada a média entre os resultados das três últimas iterações, que se mantiveram com o mesmo banco de regras e sendo executados em sumários de alta diferentes.

Tabela 5 – Resultado final da avaliação do algoritmo IRDischarge

Precisão	Revocação	Especificidade	F-Measure
93,3%	86,6%	91%	89,6%

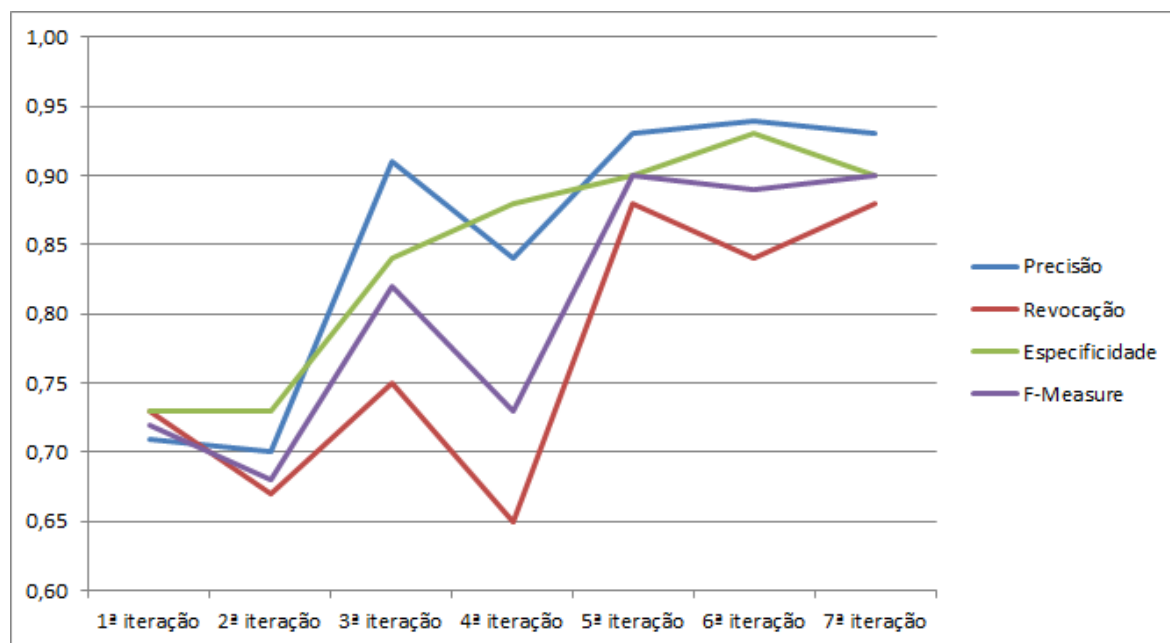


Figura 6 – Gráfico comparativo de desempenho após primeiro processamento de cada iteração

4.4 FERRAMENTA DE APOIO À ESCRITA DE SUMÁRIOS DE ALTA

Com a utilização do algoritmo IRDischarge e o banco de regras definido, desenvolveu-se uma ferramenta de apoio à escrita de sumários de alta, que alerta o profissional da saúde quanto ao preenchimento de informação referente à continuidade do cuidado no texto.

A Figura 7 apresenta um exemplo da ferramenta em funcionamento, destacando em verde a presença da continuidade do cuidado.

Preencha o sumário de alta abaixo:

Paciente HAS, dislipidêmica, cardiopata isquêmica em tratamento clínico com história familiar positiva para cardiopatia isquêmica. Interna neste hospital com quadro de ANGINA INSTÁVEL, com períodos de dor em repouso; Submetida em 06/02/04 a ACTP + colocação de STENT nas referidas lesões com resultado angiográfico satisfatório (fluxo TIMI 3). Permanece assintomática, com sinais vitais estáveis, ausência de dor torácica ou outros sinais de isquemia e com as demais comorbidades sob tratamento, paciente recebe alta com as seguintes orientações: 1- Uso correto das medicações prescritas pelo seu médico assistente (fundamental uso de AAS - Clopidogrel - Sinvastatina). 2- Retorno ao seu médico para revisão e acompanhamento do caso. 3- Retorno imediato ao seu médico ou serviço de emergência se dor torácica, febre, dor ou sangramento no local de punção, bem como outros sinais de piora do estado geral.	Identificação do paciente e caracterização da internação
	Motivo da admissão, diagnósticos e comorbidades
	Achados clínicos relevantes e outros achados
	Procedimentos diagnósticos e terapêuticos realizados
	Medicações importantes, inclusive as de alta
	Exames pendentes
	Evolução e condições de estado do paciente na alta
	Instruções relativas ao acompanhamento
	Continuidade do cuidado
	Armazenar e imprimir

Figura 7 – Ferramenta de apoio à escrita de sumários de alta com continuidade do cuidado

Para destaque da ausência da continuidade, foi utilizada a cor vermelha, conforme figura a seguir.

Preencha o sumário de alta abaixo:

Paciente interna com quadro de IAM inferior com extensão para VD, submetida a angioplastia primária com sucesso parcial. Procedimento foi complicado por possível dissecação e evento isquêmico neurológico de tronco, manifesto como paralisia de III par a D e ataxia. Ficou na UTI por 4 dias tendo recebido infusão de Dobutamina por 36 horas e ACO com heparina por 48 horas. Não apresentou angina pós-IAM, tendo boa evolução do ponto de vista hemodinâmico. Houve melhora parcial dos sintomas neurológicos. I	Identificação do paciente e caracterização da internação
	Motivo da admissão, diagnósticos e comorbidades
	Achados clínicos relevantes e outros achados
	Procedimentos diagnósticos e terapêuticos realizados
	Medicações importantes, inclusive as de alta
	Exames pendentes
	Evolução e condições de estado do paciente na alta
	Instruções relativas ao acompanhamento
Continuidade do cuidado	
Armazenar e imprimir	

Figura 8 – Ferramenta de apoio à escrita de sumários de alta sem continuidade do cuidado

5 DISCUSSÃO

Serão discutidos neste capítulo os resultados alcançados com o desenvolvimento do projeto. Inicia-se com a discussão do algoritmo de identificação de continuidade do cuidado; passando pelas regras criadas, a metodologia adotada para sua geração e o resultado de sua aplicação em um conjunto de sumários de alta; e finalizando-se com a discussão da ferramenta desenvolvida.

5.1 ALGORITMO DE IDENTIFICAÇÃO DE CONTINUIDADE DO CUIDADO, IRDISCHARGE

Um dos fatores dificultadores para a realização do processamento dos textos é a estrutura e a escrita dos mesmos, que não contém um padrão uniforme em todo banco de dados de sumários de alta. Problemas que vão desde erros ortográficos, que causam erros de etiquetação, até os diferentes padrões encontrados no texto, para escrever a mesma informação.

Vários trabalhos de pesquisa se utilizam de informações definidas em cenários ideais de estrutura e escrita (LONG, 2007; FRIEDLIN, MCDONALD, 2006). Não sendo esta a realidade dos textos utilizados neste trabalho, fazendo-se necessária, então, a fase de pré-processamento para correção de vários desses problemas.

A ausência do uso de listas terminológicas conhecidas, ou associação a palavras nas regras decorre do fato deste projeto ter seguido o modelo de regras morfológicas, como Bulegon (2011). Com a diferença que as regras não ficaram restritas a apenas três termos, porque a informação de continuidade do cuidado muitas vezes é caracterizada por trechos de texto que englobam mais de três termos, ao contrário dos diagnósticos encontrados por Bulegon.

Também a exemplo do trabalho de Bulegon, a utilização de um *corpus* clínico anotado, e a utilização da informação de gênero e número nas etiquetas morfológicas é fundamental para o sucesso do algoritmo. O primeiro, devido ao aumento considerável de acertos na etiquetação, e o segundo permitiu a geração de regras mais específicas, evitando que as regras ficassem genéricas demais e, consequentemente, levassem a mais resultados falsos positivos.

A definição dos processos de manipulação das sentenças é um diferencial

importante do algoritmo, visto que a análise apenas das últimas quatro sentenças evitou várias possibilidades de falsos positivos, e pelo levantamento feito percebeu-se que as informações referentes à continuidade do cuidado estavam nas últimas quatro sentenças em 99% dos sumários. A seleção apenas das últimas sentenças também tem um cunho pedagógico importante, pois, o correto aos profissionais é realmente informar a continuidade do cuidado como última informação do sumário de alta (SOUZA, 2012).

Outro fator que ajudou no sucesso do algoritmo é que nas últimas sentenças, além da informação de continuidade, normalmente também se encontram informações acerca das prescrições medicamentosas ao paciente. Informações estas que tem padrões bem definidos e bem distintos dos padrões morfológicos de continuidade do cuidado. Normalmente essas informações são definidas com o nome do medicamento, seguido de sua dosagem e frequência de uso. (XU, 2009).

Sistemas de RI baseados em regras, tendem a oferecer resultados confiáveis, ao mesmo tempo em que, normalmente se baseiam em regras feitas sob medida para um domínio específico, o que acaba tornando a sua reutilização mais difícil (COWIE; LEHNERT, 1996); sendo exatamente este o caso desse projeto, onde a reutilização do mesmo, depende de um novo processo de geração de regras, que é algo muito custoso para ser desenvolvido, sendo amenizado pela definição formal de uma metodologia de desenvolvimento de regras.

Abordagens que envolvam aprendizado de máquina podem oferecer soluções de mais fácil portabilidade para outros domínios, porém necessitando normalmente de uma quantidade significativa de dados para treinamento. Portanto, uma conciliação perfeita entre desempenho e portabilidade, pode ser obtida utilizando-se de uma combinação apropriada de abordagens de aprendizado de máquina e sistemas baseados em regras (SPASIC *et al.*, 2010); sendo assim, esse projeto, apesar de ser unicamente baseado em regras, sem a utilização de aprendizado de máquina, foi desenvolvido de maneira genérica o suficiente para suprir as necessidade de outros domínios, dependendo de um novo banco de regras, que poderia ser elaborado de maneira mais abrangente por meio de técnicas de treinamento de aprendizado de máquina.

5.2 BANCO DE REGRAS DO ALGORITMO

Por meio da análise dos métodos de desenvolvimento das regras, foram necessários 500 sumários de alta, de um banco de dados de 5617, para que o banco de regras final fosse construído, ou seja, 8,9% do total de sumários de alta; pois, além da etapa inicial de desenvolvimento de regras, foram necessárias mais quatro iterações para que não pudessem mais ser feitas intervenções no banco de regras para melhoria do mesmo.

A definição de que seriam também consideradas continuidade do cuidado as informações de retorno pós-alta, fez com que houvesse uma alteração significativa no banco de regras do IRDischarge, se comparado com o trabalho de Souza (2012).

O trabalho de Souza resultou em um banco de 4 regras apenas, enquanto o banco de regras atual contém 30 regras. Isso se deve ao fato das informações de retorno pós-alta apresentarem mais padrões de escrita diferentes, não ficando limitados apenas aos verbos na forma infinitiva (Ex.: manter, persistir, seguir, etc.).

Apesar de ser um método iterativo/cíclico para geração das regras, ele não pode ser considerado um método exaustivo de tentativa e erro, como os trabalhos de Bui e colaboradores (2010) e Bulegon (2011), haja vista que foram definidos critérios específicos de melhoria do banco de regras, até que os resultados se estabilizassem. Entre esses critérios estão: ordenação das regras, generalização ou especialização das regras, remoção e inserção de regras.

5.3 AVALIAÇÃO DE DESEMPENHO DO ALGORITMO

Em RI, o maior objetivo é recuperar informações relevantes dentro uma base de dados. Para medir a eficiência dos sistemas de RI são utilizados os conceitos de Precisão e Revocação (HERSH, 2009). Porém, o algoritmo desenvolvido nesta pesquisa será acoplado a uma ferramenta de apoio à escrita de sumários de alta, que tem como objetivo principal alertar ao profissional da saúde a ausência de informação acerca da continuidade do cuidado.

Portanto, se faz necessário o uso de uma métrica de RI que demonstre a eficácia do sistema ao se deparar com resultados negativos, ou seja, quando não há informação a ser recuperada, o algoritmo não buscar nenhum falso positivo. O cálculo de Especificidade é utilizado justamente para preencher essa lacuna na

avaliação dos resultados.

Dado que não foram encontradas na literatura pesquisas que identifiquem continuidade do cuidado ao paciente em sumários de alta, ou em qualquer outro tipo de narrativa clínica, é possível apenas comparar os resultados com outras pesquisas similares.

Genevieve e Hripcsak (2005) desenvolveram um método para detecção automática de eventos adversos em sumários de alta, utilizando-se de técnicas de PLN. Nesse trabalho, uma especificidade de 98% foi alcançada, e que se explica também pela baixa presença de informações requeridas no banco de dados. Dos 1000 sumários processados, apenas 32 continham eventos adversos, contribuindo para alta especificidade, que é calculada com os números de VNs e FPs.

Kovacevic e colaboradores (2013) fizeram uso do PLN para extrair eventos e expressões temporais de narrativas clínicas. Com uma combinação de regras e algoritmos de aprendizado de máquina, obtiveram como melhor marcas em seus testes de desempenho, uma precisão de 90,38% e uma revocação de 90,08%.

Em geral, os resultados obtidos nesta pesquisa na avaliação de desempenho são similares aos de outros projetos de pesquisa nas áreas de PLN e médica, como os Genevieve e Hripcsak (2005) e Kovacevic e colaboradores (2013). O ideal seria uma comparação com um projeto que identifique a mesma informação, visto que cada tipo de informação tem suas peculiaridades, como forma de escrita, frequência de aparecimento no banco de dados, etc., o que influencia no resultado final.

5.4 FERRAMENTA DE APOIO À ESCRITA DE SUMÁRIOS DE ALTA

A construção da ferramenta em uma plataforma *web* facilita seu acesso por meio de diversos dispositivos, tais como celulares, computadores, *tablets*, etc., proporcionando futura utilização dessa ferramenta pelo profissional da saúde, em ambientes hospitalares, da maneira mais conveniente e oportuna.

O desenvolvimento da ferramenta foi baseado no conceito de acoplamento, para que futuramente outros algoritmos que identifiquem outros elementos-chave essenciais ao sumário de alta, alertando ao profissional que elaborar o sumário, quanto à ausência de quaisquer informações vitais para qualidade de atendimento ao paciente, como diagnósticos, procedimentos realizados e medicações aplicadas (SOUZA, 2012).

Os sistemas de apoio à decisão em saúde têm por objetivo: auxiliar profissionais da saúde a tomarem decisões clínicas, lidar com dados médicos dos pacientes ou com o conhecimento médico necessário para interpretação dos dados. (SHORTLIFFE, 1987). Pode-se afirmar, portanto, que a ferramenta gerada a partir desse projeto, pode ser caracterizada como uma ferramenta de apoio à decisão em saúde, pois auxilia diretamente o profissional de saúde a lidar com dados médicos do paciente, enquanto ele descreve um sumário de alta.

A exemplo da ferramenta desenvolvida por Stetson e colaboradores (STETSON *et al.* 2008), a ferramenta construída neste projeto de pesquisa visa a melhoria da qualidade dos sumários de alta escritos em ambientes computacionais, impactando de maneira positiva a qualidade do cuidado ao paciente e segurança do mesmo.

5.5 TRABALHOS FUTUROS

Por meio do uso da ferramenta de apoio à escrita de sumários de alta, pode-se replicar ou desenvolver algoritmos para identificar todos os elementos-chave essenciais dentro do conteúdo dos sumários de alta. Para que o profissional da saúde seja alertado quanto à ausência de informações importantes, e que as mesmas sejam preenchidas corretamente, podendo assim, impactar diretamente na qualidade de atendimento ao paciente em ambientes hospitalares.

A reutilização do algoritmo para identificação de continuidade do cuidado em outros domínios, ou até mesmo para identificação de outras informações presentes no domínio médico, depende de um novo banco de regras para o algoritmo. O desenvolvimento de um método semi-automatizado de geração de regras (treinamento do algoritmo), utilizando-se de conceitos de aprendizado de máquina podem suprir outras necessidades de identificação de informação, bem como melhorar os resultados atuais do algoritmo, pois o mesmo passará por um processo mais extensivo de treinamento. Atualmente todo esse processo é feito de maneira manual, exigindo maior tempo de desenvolvimento.

Uma interface de treinamento do sistema voltada ao usuário final (profissionais da saúde) pode auxiliar a ferramenta de apoio à escrita a se adequar aos mais diferentes padrões de escrita, de diferentes instituições médicas. Tornando-a assim mais maleável, e sem necessidade de um profissional com

conhecimento técnico do sistema para intervir na utilização da ferramenta.

6 CONCLUSÃO

O algoritmo IRDischarge, juntamente com o banco de regras desenvolvido, mostrou-se funcional na habilidade de identificação da continuidade do cuidado ao paciente, e obteve resultado satisfatório, mesmo sendo executado em um banco dados onde os sumários não estavam em uma estrutura e escrita ideais.

A ferramenta de apoio à escrita, combinada com algoritmos de identificação de elementos essenciais ao sumário de alta, pode auxiliar na elaboração e, conseqüentemente, no aumento de qualidade dos sumários de alta desenvolvidos em ambientes hospitalares, o que pode aumentar a qualidade de atendimento ao paciente. E com a utilização de um corretor ortográfico, a ferramenta pode inclusive evitar erros nos resultados do algoritmo, provindos de erros de digitação por parte do profissional da saúde.

A utilização do algoritmo em outros domínios da saúde também é um ponto relevante do projeto, pois, existe uma metodologia bem definida para geração de regras que, se executada novamente, pode gerar um conjunto de regras para identificar outros tipos de informação em narrativas clínicas. É importante também ressaltar que o processo que se utiliza apenas das quatro últimas sentenças deve ser retirado do algoritmo se houver necessidade de buscar outras informações, pois, o mesmo refere-se a uma característica específica da continuidade do cuidado.

Este projeto de pesquisa pode ser inserido no dia-a-dia hospitalar, e inclusive pode ser incorporado a sistemas de Prontuário Eletrônico do Paciente, durante o preenchimento dos sumários de alta.

REFERÊNCIAS

ARAMPATZIS, AT; VAN DER WEIDE, TP; VAN BOMMEL, P; KOSTER CHA. Linguistically-motivated Information Retrieval. Encyclopedia of Library and Information Science, v.69, p.201-222, 1999.

BAEZA-YATES, R; RIBEIRO-NETO, B. Modern Information Retrieval: The Concepts and Technology Behind Search. Second Edition. Addison Wesley, 2011.

BITENCOURT, JL. Monitoração de procedimentos em um tesauro multilíngue. 2006. Dissertação (Mestrado em Tecnologia em Saúde) – Pontifícia Universidade Católica do Paraná, Curitiba, 2006.

BUI, QC; NUALLÁIN, BO; BOUCHER, CA; SLOOT, PM. Extracting causal relations on HIV drug resistance from literature. BMC Bioinformatics. 2010 Feb.;11(101):11-23.

BULEGON, H; MORO, CMC; GIRARDI, DR. NOTTUS – Sistema para Identificação de Diagnósticos Aplicando Processamento de Linguagem Natural. In: XII Congresso Brasileiro de Informática em Saúde, 2010, Porto de Galinhas - PE. Anais do XII Congresso Brasileiro de Informática em Saúde, 2010.

BULEGON, H. Identificação de diagnósticos contidos em narrativas clínicas e mapeamento para a classificação internacional de doenças. 2011. 101f. Dissertação (Mestrado em Tecnologia em Saúde) – Pontifícia Universidade Católica do Paraná, Curitiba, 2011. Disponível em: <http://www.biblioteca.pucpr.br/tede/tde_busca/arquivo.php?codArquivo=1861>. Acesso em: 1 de março 2013.

CHEN, ES; HRIPCSAK, G; FRIEDMAN, C. Disseminating Natural Language Processed Clinical Narratives. AMIA 2006 Symposium Proceedings. p. 126. 2006.

CHIANG, JH; LIN, JW; YANG, CW. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language

Extraction and Encoding System (MedLEE). Journal American Medical Informatics Association 2010;17:245-252.

CHOWDHURY, GG. Natural language processing. Annual Review of Information Science and Technology. n 37, p. 51-89. 2003.

CoGrOO. Corretor gramatical acoplável ao OpenOffice.org. Disponível em: <<http://cogroo.sourceforge.net>>. Acesso em: 1 de março de 2013.

COPPIN, B; Inteligência artificial. Tradução e revisão técnica Jorge Duarte Pires Valério. Rio de Janeiro: LTC, 2010.

COWIE J, LEHNERT W. Information extraction. Commun ACM 1996;39:80-91.

CRAIG, J; CALLEN J; MARKS A; SADDIK, B; BRAMLEY, M. Electronic discharge summaries: the current state of play. Health Information Management Journal. Vol. 36 n. 3. 2007.

EMBI, PJ; YACKEL, TR; LOGAN, JR; BOWEN, JL; COONEY, TG; GORMAN, PN. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. Journal American Medical Informatics Association. 2004;11:300–309.

ERTMER, A; UCKERT, F. User acceptance of and satisfaction with a personal electronic health record. Studies In Health Technology And Informatics. Vol. 116, p. 861-866. 2005.

FRIEDLIN, J; MCDONALD, CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annual Symposium Proceedings 2006; p. 925, 2006.

FRIEDMAN, C; ALDERSON, PO; AUSTIN, JHM; CIMINO, JJ; JOHNSON SB. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association. Vol. 1, n.2, p.161-174, mar./abr. 1994.

FRIEDMAN, C; SHAGINA, L. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Proc AMIA Symp.* 1999:256-60.

FRIEDMAN, C; HRIPCSAK, G; DUMOUCHEL, W; JOHNSON, SB; CLAYTON, PD. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1, p. 83-108. 1995.

GALVÃO, MCB; RICARTE, ILM. O prontuário eletrônico do paciente no século XXI: contribuições necessárias da ciência da informação. *Ver. Inf. E Doc. Ribeirão Preto*, v.2, n.2, p.77-100, jul./dez. 2011.

GATE – General Architecture for Text Engineering. Disponível em: <<http://gate.ac.uk/>>. Acesso em: 1 de março de 2013.

GENEVIEVE, MB; HRIPCSAK, G. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. *Journal American Medical Informatics Association*. 2005;12:448-457.

GONNELA, JS; HERMAN, MW. Continuity of care. *JAMA*. 1980;243(4):352-354.

GONZALEZ, M; LIMA, VLS; LIMA JV. Termos, relacionamentos e representatividade na indexação de texto para recuperação de informação. *Letras de Hoje. Porto Alegre*. v. 41, n. 2, p. 65-87, junho/2006.

GUNDLAPALLI, AV; SOUTH, BR; PHANSALKAR, S; KINNEY, AY; SHEN S; DELISLE S; PERL, T; SAMORE, MH. Application of Natural Language Processing to VA Electronic Health Records to Identify Phenotypic Characteristics for Clinical and Research Purposes. *Translational Bioinformatics*. 2008; 2008: 36–40.

HAGGERTY, JL; REID, RJ; FREEMAN, GK; STARFIELD, BH; ADAIR, RM. Continuity of care: a multidisciplinary review. *BMJ*, Vol. 327, p. 1219-1221. Nov/2006.

HERSH, W. *Information Retrieval: A Health and Biomedical Perspective*. Third

Edition. Springer, 2009.

IEEEXplore – Digital Library. Disponível em: < <http://ieeexplore.ieee.org/>>. Acesso em: 1 de março de 2013.

ISO - Health informatics - Electronic health record - Definition, scope and context. Disponível em: < http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39525>. Acesso em: 1 de março de 2013.

JURAFSKY, D; MARTIN, J. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, Pearson Prentice Hall. Stanford University, New Jersey. 2008.

KERGOAT, MJ; LATOUR, J; JULIEN, I; PLANTE, MA; LEBEL, P; MAINVILLE, D; BOLDUC, A; BUCKLAND, JA. A discharge summary adapted to the frail elderly to ensure transfer of relevant information from the hospital to community settings: a model. BMC Geriatrics 2010, 10:69.

KOVAČEVIĆ A, DEHGHAN A, FILANNINO M, KEANE JA, NENADIC G. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. Journal American Medical Informatics Association; 2013; 20:859-866.

KRIPALANI, S; LEFEVRE F; PHILLIPS, CO; WILLIAMS, MV; BASAVIAH, P; BAKER, DW. Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. JAMA, 28 fev. Vol. 297, n.8. 2007.

LAPELLE, NR; LUCKMANN, R; SIMPSON, EH; MARTIN ER. Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. BMC Public Health. 2006; 6:89.

LONG W. Lessons extracting diseases from discharge summaries. AMIA Annual Symposium Proceedings. 2007; p. 478–482, 2007.

MATHENY, ME; FITZHENRY, F; SPEROFF, T; GREEN, JK; GRIFFITH, ML; VASILEVSKIS, EE; FIELSTEIN, EM; ELKIN, PL; BROWN, SH. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. International Journal of Medical Informatics. Vol. 81, p. 143-156, 2012.

MANARIS, B. Natural Language Processing: A Human-Computer Interaction Perspective. Advances in Computers (Marvin V. Zelkowitz, ed.), vol. 47, p. 1-66, Academic Press, New York, 1998.

MANNING, CD; SCHÜTZE, H. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: Maio 1999.

MANNING, CD; RAGHAVAN, P; SCHÜTZE, H. Introduction to Information Retrieval, Cambridge University Press. 2008.

MARTHA, AS; CAMPOS, CJR; Sigulem, D. Recuperação de Informações em Campos de Texto Livres de Prontuários Eletrônicos do Paciente Baseada em Semelhança Semântica e Ortográfica. J. Health Inform. Jul-Set; 2(3): 63-71, 2010.

MASSAD, E; MARIN, HF; NETO, RSA. O prontuário eletrônico do paciente na assistência, informação e conhecimento médico. São Paulo: p.213, 2003.

MENEZES, C; NETO, J. Um método híbrido para a construção de etiquetadores morfológicos, aplicado à língua portuguesa, baseado em autômatos adaptativos. Anais da Conferencia Iberoamericana en Sistemas, Cibernética e Informática, 19-21 de Julho, 2002, Orlando, Florida.

MEYSTRE, SM; SAVOVA, GK; KIPPER-SCHULER, KC; HURDLE, JF. Extracting Information from Textual Documents in the Electronic Health Record - A Review of Recent Research. IMIA Yearbook of Medical Informatics 2008. Methods Inf Med

2008; 47 Suppl 1:128-44

MOORE, C; WISNIVESKY J; WILLIAMS, S; MCGINN, T. Medical errors related to Discontinuity of care from an inpatient to an outpatient setting. *Journal of General Internal Medicine*, 2003;18:646-651.

MYERS, JS; JAIPAUL, K; KOGAN, JR; KREKUM, S; BELLINI, LM; SHEA, JA. Are Discharge Summaries Teachable? The Effects of a Discharge Summary Curriculum on the Quality of Discharge Summaries in an Internal Medicine Residency Program. *Academic Medicine*, Vol. 81, N. 10. Outubro/2006.

NADKARNI, PM; OHNO-MACHADO, L; CHAPMAN, W. Natural Language Processing: an introduction. *Journal American Medical Informatics Association*; 2011;18:544-551.

NLTK – Natural Language Toolkit. Disponível em: <<http://nltk.org/>>. Acesso em: 1 de março de 2013.

OpenNLP – Apache OpenNLP library for processing of natural language text. Disponível em: <<http://opennlp.apache.org/>>. Acesso em: 1 de março de 2013.

ORENGO, VM; HUYCK, C. A Stemming algorithm for the Portuguese Language. In *Proceedings of SPIRE'2001 Symposium on String Processing and Information Retrieval*, Laguna de San Raphael, Chile. Disponível em: <<http://www.computer.org/csdl/proceedings/spire/2001/1192/00/11920186-abs.html>>. Acesso em: 1 de março de 2013.

PACHECO EJ. MorphoMap: mapeamento automático de narrativas clínicas para uma terminologia médica. [Tese]. Curitiba: Universidade Tecnologia Federal do Paraná; 66 2009. 154 p. Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial.

PERREN, A; PREVISDOMINI, M; CERUTTI, B; SOLDINI, D; DONGUI, D; MARONE, C. Omitted and unjustified medications in the discharge summary. *Quality Safe*

Health Care 2009;18:205–208.

PETERS, AC; OLEYNIK, M; PACHECO, EJ; MORO, CMC; SCHULZ, S; NOHAMA, P. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. In: XII Congresso Brasileiro de Informática em Saúde, 2010, Porto de Galinhas - PE. Anais do XII Congresso Brasileiro de Informática em Saúde, 2010. p. p1-p6.

PINTO, VB; Prontuário eletrônico do paciente: documento técnico de informação e comunicação do domínio da saúde. Revista eletrônica biblioteconomia. Florianópolis. n.21, 1º sem. 2006.

PUBMED - Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/>>. Acesso em: 1 de março de 2013.

ROUGHEAD, EE; KALISCH, LM; RAMSEY, EN; RYAN, P; GILBERT, AL. Continuity of care - when do patients visit community healthcare providers after leaving hospital. Internal Medicine Journal, 2011 Setembro;41(9):662-7

SCIENCEDIRECT – scientific database. Disponível em: <<http://www.sciencedirect.com/>> Acesso em: 1 de março de 2013.

SHEN, S; SOUTH, BR; BUTLER, J; BARRUS, R; WEIR, C. The relationship between structural characteristics of 2010 challenge documents and ratings of document quality. AMIA Annual Symposium Proceedings. 2012;2012:848-55.

SHORTLIFFE, EH; BARNETT GO. Medical data: their acquisition, storage, and use. 2a. ed. Nova Iorque: Springer. 2001.

SHORTLIFFE EH. Computer programs to support clinical decision making. JAMA 1987;258(1):61–6.

SOUZA, AC. Identificação Do Conteúdo Padronizado Do Sumário De Alta. 2012. Dissertação (Mestrado em Tecnologia em Saúde) – Pontifícia Universidade Católica

do Paraná, Curitiba, 2012.

SPASIC, I; SARAFRAZ, F; KEANE JA; NENADIC G. Medication information extraction with linguistic pattern matching and semantic rules. Journal American Medical Informatics Association, 2010;17:532-535.

STETSON, PD; MORRISON, FP; BAKKEN, S; JOHNSON, SB. Preliminary development of the physician documentation quality instrument. Journal American Medical Informatics Association, 2008 Jul-Aug;15(4):534-41.

UZUNER, Ö; SIBANDA TC; LUO Y; SZOLOVITS, P. A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine. v.42, n.1, p.13-35. 2008.

VAN WALRAVEN, C; WEINBERG, AL. Quality assessment: of a discharge summary system. Can Med. Assoc. J. Maio Vol.1, p. 152 (9), 1995.

VAN WALRAVEN, C; ROKOSH, E. What Is Necessary for High-Quality Discharge Summaries? American Journal of Medical Quality Vol.14, p.160. 1999.

VAN WALRAVEN, C; SETH, R; LAUPACIS, A. Dissemination of discharge summaries - Not reaching follow-up physicians. Canadian Family Physician Vol. 48, p. 737-742. Abril/2002.

VISL Portuguese Grammar. Disponível em: <<http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html>>. Acesso em: 1 de março de 2013.

WANG, X; HRIPCSAK, G; MARKATOU, M; FRIEDMAN, C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal American Medical Informatics Association. 2009;16:328 –337.

WU, Y; DENNY, JC; ROSENBLOOM, ST; MILLER, RA; GIUSE, DA; XU, H. A comparative study of current clinical natural language processing systems on

handling abbreviations in discharge summaries. AMIA Annual Symposium Proceedings. 2012;2012:997-1003

XU, H; STENNER, SP; DOAN, S; JOHNSON, KB; WAITMAN, LR; DENNY, JC. MedEx: a medication information extraction system for clinical narratives. Journal American Medical Informatics Association 2010;17:19-24.

ZENG, QT; GORYACHEV, S; WEISS, S; SORDO, M; MURPHY, SN; LAZARUS, R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Informatics and Decision Making 2006, 6:30.