Processamento de Linguagem Natural aplicado à anamneses do domínio da ginecologia

Amanda Damasceno de Souza[1]

Eduardo Ribeiro Felipe[2]

Resumo: O volume de informação produzido tanto na pesquisa médica quanto na prática clínica há muito exige tratamento computacional. Uma importante fonte de dados reais, relevante para a pesquisa, além de essencial para a gestão das unidades de saúde, é o prontuário de paciente. Dessa forma, técnicas de processamento de linguagem natural (PLN) são alternativas importantes para lidar com essa fonte dinâmica onde constantemente se registram novos dados. A presente pesquisa se insere nesse contexto, exibindo uma iniciativa de extração de dados de pacientes em um grande hospital através de técnicas de PLN. Para tanto, apresenta-se um *background* com informações sobre tais técnicas para então descrever os passos metodológicos, bem como os resultados parciais da extração realizado no campo médico da ginecologia.

Palavras-chave: Registros Eletrônicos de Saúde; Anamnese; Terminologia; Processamento de Linguagem Natural.

Natural Language Processing applied to anamnesis in the field of gynecology

Abstract: The volume of information produced both in medical research and in clinical practice have required, since much time, the application of computational resources. An important source of real data, which is relevant for research, in addition to being essential for the management of health units, is the set of healthcare patient records of a hospital. Natural language processing (PLN) techniques consist of an important alternative to address this dynamic source in which new data is constantly recorded. The present research is conduct in this context, exhibiting an initiative to extract data from patient's records in a large hospital using PLN techniques. To reach our goals, we present background information about such techniques, in adddition to methodological steps and partial results of the extraction conducted in the field of gynecology.

Keywords: Electronic Health Records; Medical History Taking; Terminology; Natural Language Processing.

[1] Doutora em Gestão e Organização do Conhecimento - Universidade Federal de Minas Gerais, Bibliotecária do Núcleo de Ciências da Saúde Felício Rocho/Hospital Felício Rocho. e-mail: amandasd81@gmail.com Lattes: http://lattes.cnpq.br/3615797323442040. ORCID: https://orcid.org/0000-0001-6859-4333.

[2] Doutor em Gestão e Organização do Conhecimento da Informação - Universidade Federal de Minas Gerais. Professor Adjunto no curso de Engenharia de Computação na Universidade Federal de Itajubá (UNIFEI) Campus Itabira, e-mail: eduardo.felipe@unifei.edu.br Lattes: http://lattes.cnpq.br/1010588591399870. ORCID: https://orcid.org/0000-0003-1690-2044.

1 INTRODUÇÃO

No âmbito da saúde, o Prontuário Eletrônico do Paciente (PEP) é uma fonte importante de dados em saúde, mas o fato da maioria de seus dados serem registrados de forma não padronizada – dados não estruturados ou semi estruturados – dificulta a utilização destes no processo de Recuperação e em pesquisas científicas (Wang et al, 2012). Os PEPs são ricos em informação no campo de anamnese, a qual, em grande parte, se apresenta na forma de texto livre. Meios para extrair informação de registros em texto livre exigem um significativo esforço de pesquisa (Zhou et al, 2006).

A palavra "anamnese" é originada do grego *anamnesis* e diz respeito a reminiscência, ao ato de relembrar. No contexto da medicina diz respeito ao registro completo da história clínica de um paciente (Farlex, 2012). A elaboração da anamnese e a realização de exame clínico são funções primordiais dos médicos para relatar problemas de saúde dos pacientes. Segundo López (1990), a anamnese é "essencial para a prática da medicina integral, isto é, da medicina que se preocupa com os aspectos biopsicossociais das moléstias" (p.20). Através da anamnese é recolhida informação sobre fatos de interesse médico a respeito da vida dos pacientes pois trata-se de um método de diagnóstico (Lopez, 1990). O diagnóstico assertivo e a comunicação entre a equipe de saúde dependem da avaliação clínica e o encaminhamento correto da informação proveniente da anamnese (Grüne, 2016). No Brasil, a elaboração de anamnese em Prontuário do Paciente é obrigatória, de acordo com a Resolução CFM nº 2056, de 20 de setembro de 2013 e com o Código de Ética Médica publicado pela Resolução CFM nº 1.931, de 17 de setembro de 2009 (Brasil,2013; CFM, 2009).

Neste contexto, o processamento de linguagem natural (PLN) tem potencial para contribuir com soluções no âmbito da extração e estruturação de informação clínicas textuais para disponibilizar dados clínicos para uso na tomada de decisão (Friedman e Hripcsak, 1999). Dados e termos provenientes de textos clínicos de anamnese podem auxiliar a tomada de decisão em saúde, a pesquisa científica e também a criação e enriquecimento de terminologias clínicas. Por "terminologias clínicas", entendem-se aqui um conjunto de artefatos para fins de representação que inclui "classificação", "tesauro", "vocabulário", "nomenclatura" e "ontologia"

(Schulz et al., 2017). A extração de termos de narrativas clínicas por meio de PLN é importante e auxilia aos desenvolvedores de ontologias na identificação de conceitos relevantes no domínio da medicina (Baneyx, Charlet e Jaulent, 2006). A extração automática de vocabulários provenientes de narrativas médicas de PEPs pode ajudar a melhorar e manter terminologias clínicas, como SNOMED e as ontologias biomédicas (Spackman e Hersh, 1996). Com a extração de termos a partir de texto livre é possível também verificar quais assuntos são abordados na prática clínica, em certo momento.

A recuperação da informação e do conhecimento presente nos textos livres em linguagem natural é uma tarefa árdua que atualmente demanda técnicas de PLN. O PLN diz respeito ao conjunto de técnicas para processamento de texto em linguagem natural, que se utiliza de métodos da linguística computacional (Manning e Schütze, 1999). Envolve técnicas como a mineração de texto e utiliza conhecimento multidisciplinar da Linguística, da Linguística Computacional, das Ciências da Computação, da Inteligência Artificial, da Matemática, da Lógica, da Filosofia, da Estatística e da Psicologia para realizar a análise da linguagem humana, dentre outras possibilidades úteis (Dias-Da-Silva, 2006).

A presente pesquisa, conduzida no âmbito da Ciência da Informação (CI), é uma iniciativa de aplicação de PLN com vistas a recuperação da informação de anamneses de prontuários eletrônico do campo médico da ginecologia. A CI, nas tarefas de organização do conhecimento, atua em domínios como da Medicina buscando soluções para os problemas de informação e para a melhor gestão dos recursos em saúde (Ciol e Beraquet, 2009). De fato, o processamento de linguagem natural, a linguística computacional, aspectos da inteligência artificial e também as áreas de *text mining, web mining* e *data mining*, estão entre as técnicas que a CI se vale para conduzir sua pesquisa (Almeida, Souza e Baracho, 2015).

O objeto de estudo da presente pesquisa é o Prontuário Eletrônico do Paciente (PEP) do Hospital Felício Rocho (HFR), um grande hospital de Belo Horizonte, onde o trabalho cientifico foi aprovado pelo Comitê de Ética em Pesquisa (CEP), sob o número CAAE:03384418.0.0000.5125. O objetivo da pesquisa é a aplicação de técnicas de mineração de textos (*Text Mining*) em anamneses de prontuário

eletrônico do paciente para extração de termos visando enriquecimento de terminologia clínica do tipo ontologia.

2 REVISÃO DE LITERATURA

Com o desenvolvimento crescente das tecnologias de informação, uma equipe médica produz hoje uma quantidade de informação maior do que em qualquer outro momento da história. Grande parte desta informação está em formato texto e digital. A sobrecarga de informação resultante de tanto material disponível impacta na tomada de decisão, sendo necessário utilizar recursos tecnológicos para recuperar conteúdo relevante e que possa ser interoperável com as terminologias clínicas. Neste contexto, a recuperação de informação é entendida como um conjunto de abordagens para análise de conteúdo em linguagem natural, incluindo o processamento de linguagem natural e a mineração de texto ou *text mining* (TM).

2.1 Terminologias, ontologias e text mining

Na literatura, a palavra "terminologia" apresenta três significados principais (Campos, 2001): a) Uma lista de termos e seus significados; b) Os termos de uma área de especialidade; c) Um conjunto de princípios teóricos. O primeiro significado é relacionado aos dicionários, vocabulários e léxicos, se referindo a uma apresentação ordenada de conceitos; já o segundo se refere ao campo de estudo científico dos termos de uma área especializada; o terceiro é aquele relativo ao campo de estudo teórico da terminologia e se refere a um campo do saber, a disciplina terminologia (Campos, 2001).

As terminologias são utilizadas com os objetivos principais de: a) apoiar o software clínico, para construir PEPs e sistemas de apoio à decisão assistida por computador, com garantia de qualidade e de gestão de informação; b) apoiar a conversão de esquemas de codificação epidemiológica e de relatórios existentes, tais como CID 9/10; c) fomentar o intercâmbio multilíngue, por estar disponível na língua dos profissionais da saúde que as utilizam.

Há uma diversidade de terminologia clínicas com diferentes propósitos, por exemplo: as que representam o jargão médico e são denominadas "terminologias

de interface"; as ontologias, que lidam com o conhecimento canônico, muitas vezes rotuladas de "terminologias de referência"; e as classificações, como CID-10, chamadas de "terminologias de agregação" (Schulz et al., 2017). Entres as diversas terminologias clínicas, as ontologias vêm ganhando destaque na área de saúde em meio a crescente necessidade de gerenciamento inteligente de informação e conhecimento com vistas a interoperabilidade de conteúdo (Freitas, Schulz e Moraes, 2009).

Com relação aos procedimentos de "*mining*", o processamento de textos clínicos e biomédicos no âmbito da informática médica envolve a utilização de métodos baseados em PLN, a qual contempla técnicas como o *Text Mining* (TM) ou mineração de texto. O sistema de mineração de texto objetiva identificar padrões significativos e "aprender" sobre o espaço de informação relacionado à necessidade de recuperação (Blake, 2011). O PLN envolve processamento inteligente de texto, no qual o computador buscar interpretar o que foi escrito em linguagem natural, valendo-se de métodos computacionais linguísticos. Essas duas abordagens, de TM e PLN, visam a extração de informação específica de documentos ou coleções de documentos, de forma que podem ser aplicadas em campos de texto livre dos PEPs (Dalianis, 2018).

2.2 Breve visão dos trabalhos relacionados

Pesquisas envolvendo mineração de texto clínico – ou *Text Mining* – para extração de informação de texto clínico de PEPs têm sido realizadas há muito. No estudo de Kim et al. (2012) utilizou-se o TM para extrair informação sobre intervenção coronária percutânea. Já Wang et al. (2012) utilizou técnica de *Machine Learning* por meio do *Support Vector Machine* (SVM) para extrair resultados de diagnóstico de textos clínicos do PEP sobre angiografia coronariana e câncer de ovário. O estudo de Zhou et al. (2019) descreveu um sistema de extração de informação médica, para extrair uma variedade de informação e registros clínicos de textos clínicos do paciente sobre queixas de doenças da mama. Já Meystre et al. (2010) fizeram uma revisão da literatura sobre pesquisas recentes em desidentificação de documentos de texto clínico narrativo em PEP. Estes estudos relatam as possibilidades, bem como a importância de se realizar PNL em campos abertos do PEP.

3 METODOLOGIA: TEXT MINING NA EXTRAÇÃO DE DADOS

Para a metodologia desta pesquisa, optou-se em utilizar técnicas de TM e PLN direcionadas a recuperação da informação de texto. Dentre os métodos mais comuns usados para analisar textos citam-se (Black, 2011; Wallach, 2006; Barion, 2008).

3.1 Principais métodos

a) Recursos no nível da superfície: capturam informação sobre palavras ao identificar características da própria palavra, por exemplo, nomes próprios de cidades, pessoas e organizações são reconhecidos e diferenciados de outras palavras por começarem com letra maiúscula. O outro exemplo é o caso dos genes, em que a identificação por um recurso no nível da superfície ocorre por meio de inferência de que tais nomes podem incluir números romanos ou mistura de letras maiúsculas, minúsculas e números.

b) Representação baseada em vetores: a expressão *Bag of Words* (BOW), traduzida literalmente como "saco de palavras", é um recurso muito utilizado no TM. Trata-se uma abordagem em que o sistema representa cada documento como um vetor ponderado de termos, e o peso associado a cada termo é o número de vezes que ele aparece no documento. Nesse caso, não é necessário conhecimento de domínio, utilizam-se métodos de análise de similaridade entre documentos, como por exemplo o *clustering*[3]. Uma questão relevante, porém, é como definir um termo: em sistemas no idioma inglês, um termo é definido pelo "conjunto contínuo de caracteres alfanuméricos que ocorre entre espaços em branco e pontuação".

c) Representação de conceito: para uma boa representação dos textos, problemas como sinonímia (quando palavras diferentes possuem o mesmo significado) e polissemia (quando a mesma palavra possuir significados diferentes) devem ser solucionados. Para tal solução, é recomendável a utilização de termos e conceitos sobre esses termos representados em artefatos ontológicos, em uma terminologia padronizada que faça uso de teorias da ontologia, para que um

[3] Significa **análise de agrupamento de dados,** realiza agrupamentos automáticos de dados segundo o seu grau de semelhança. Fonte: https://pt.wikipedia.org/wiki/Clustering. 2019.

termo seja representado uma única vez, de maneira formal, evitando-se ambiguidade (Almeida, 2020).

d) Análise de n-gramas, bi-gramas, etc: analisa a frequência das expressões, ou seja, faz previsões usando frequências marginais e condicionais de palavras observadas no texto.

e) Extração: identifica no texto, de forma precisa, conjuntos pré-definidos de termos representativos de entidades: nomes, organizações, locais, proteínas, genes, datas, horas, valores monetários, porcentagens, etc. Esta atividade é usada no PLN para manipular e transformar dados não estruturados na descoberta de conhecimento.

Ao elaborar estratégias de extração de informação é necessário ter claramente registrado o que se pretende buscar. Enfatiza-se isso porque, uma das dificuldades com abordagens de aprendizagem é a necessidade de exemplos de treinamento. Isso quer dizer que as abordagens ou algoritmos utilizados no PLN precisam ser previamente treinados para que seja estabelecido que tipo de informação se pretende buscar e recuperar (Blake, 2011).

3.2 Estratégias

Ao elaborar estratégias de extração de informação utilizando as abordagens de PLN, é necessário analisar os seguintes aspectos:

1. Segmentação e Tokenização: é uma das primeiras etapas do PLN, a qual realiza a tarefa de separar as sentenças e as palavras (Dalianis, 2018). Permite detectar os limites de “token” e partes do discurso, ou seja, uma palavra que será analisada nas tarefas subsequentes do processamento morfológico (Dalianis, 2019, IBM, 2018). Em idiomas como inglês, português, espanhol, por exemplo, os tokens são identificados por espaços típicos da sintaxe.

2. *Case Folding*[4]: é o processo de padronização do texto na forma de uma representação que envolva apenas caracteres minúsculos, para propósitos de comparação.

3. Processamento morfológico: faz análise de artigos, verbos, substantivos e adjetivos no texto. (Santos et al., 2015). Este recurso envolve: *Stemming e Compound splitting* – a) *Stemming*: busca-se identificar a raiz do termo, reduzindo a palavra à sua raiz, sem considerar a classe gramatical (Dalianis, 2018). Por exemplo: *differ, different, differing* e *differs*, seriam todos representados como *differ* ao se aplicar o *stemming* (Blake 2011). Exemplo em português: amigo, amiga, amigão, seriam representados por amig; e os termos: gato, gata, gatos, gatas, seriam representados por: gat. O *stemming* é um recurso importante para realizar a remoção de variações de palavras. As variações são identificadas pelos prefixos e sufixos, assim como gerúndios e plurais (Barion, 2008). b) *Compound splitting*: a divisão de compostos consiste em decompor palavras estrangeiras, uma vez que a composição de linguagem é comum em certos idiomas como, por exemplo, alemão e sueco (Dalianis, 2018).

4. Abreviaturas: uma abordagem importante em sistemas médicos é a análise de abreviaturas para identificar as abreviações adotadas na literatura e identificar a sua forma expandida correta (Blake 2011). Além disso, é importante verificar a presença de acrônimos, verificara ortografia, a corrigir erros, marcar partes de fala (Dalianis, 2018). No domínio da ginecologia, os especialistas utilizam abreviações de procedimentos cirúrgicos nos campos de texto livre de um PEP, por exemplo: histerec para histerectomia, ooforec para ooforectomia, episio para episiotomia, vulvec para vulvectomia, bartolinec para bartholinectomia, miomec para miomectomia. Além das abreviações dos termos, também é usual encontrar siglas para os procedimentos realizados, por exemplo: HAT significa "Histerectomia Abdominal Total".

5. Análise sintática de negação: a análise sintática apresenta enfoque na captura de informação no nível da sentença para interpretar frases com palavras idênticas, mas com sentidos diferentes. Na verificação de negação, busca-se analisar a

[4] https://www.w3.org/TR/charmod-norm/#definitionCaseFolding

presença de sentença de negação nos textos médicos, tarefa importante devido à presença de resultados de testes negativos, que são resultados de testes anormais em contraste com testes anteriores.

6. Redução de dimensionalidade: ao reduzir a dimensionalidade pode-se reduzir o ruído na coleta de texto original e, assim, fornecer padrões (Blake 2011).

7. Extração de conceitos e de relações: na análise semântica é realizada a tarefa de interpretar os significados ou identificar as entidades semânticas, processo também chamado de análise de texto. Para realizar a análise semântica são empregadas várias técnicas, dentre elas o reconhecimento de entidades, a detecção de negação, a extração de relações, para citar algumas. O *Named Entity Recognition* (NER), ou reconhecimento de entidades, visa identificar nomes pessoais, localidades, organizações, datas etc., bem como outras entidades de interesse no texto como sintomas, doenças, medicamentos e partes do corpo (Dalianis, 2018). Nessa etapa de análise semântica, a ontologia biomédica é um recurso fundamental, pois, após utilizar as ontologias para identificar conceitos representativos de entidades em fragmentos textuais, identificam-se as relações entre essas entidades (Dalianis, 2018).

3.2 Os passos metodológicos

A amostra da pesquisa foi composta por anamneses de pacientes atendidas na instituição na clínica de ginecologia, origem do atendimento (ambulatório e internação), no ano de 2018. Além dos textos das anamneses, foram utilizadas as variáveis CID/10 e faixa etária. Para a coleta de dados nos PEPs do HFR foi planejada uma estratégia inspirada em *Business Intelligence* (BI), com a finalidade de recuperar somente os dados de interesse e preparar o banco de dados para a realização do *Text Mining*. A estratégia de BI foi importante para identificar quais campos de anamnese em texto livre a equipe de ginecologia fazia uso para preencher os dados no sistema do hospital (o MV-PEP).

Como resultado, um banco de dados relacional em PostgreSQL foi exportado a partir do processo de identificação dos registros de interesse do projeto, no banco de

dados do hospital. Os processos e análise dos dados são descritas a seguir. A figura 1 permite verificar de forma gráfica, as etapas principais:

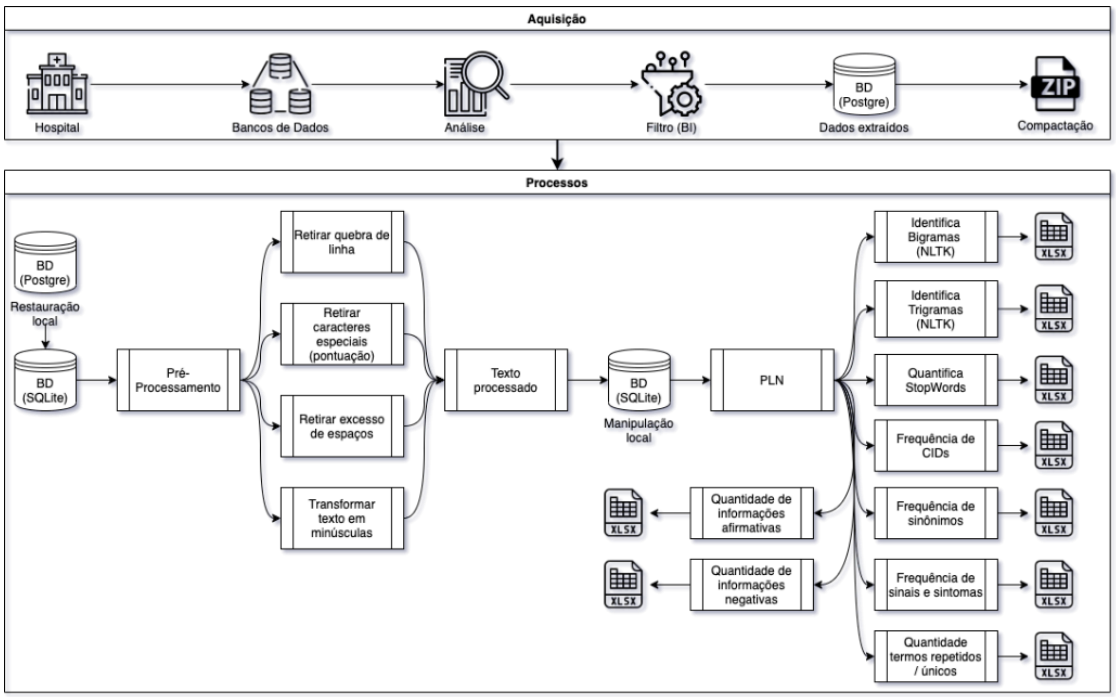

Figura 1 - Processos e análise dos dados da pesquisa
Fonte: Dados da pesquisa, 2020.

Como processos iniciais, destacam-se: a) a extração da informação do banco de dados principal do hospital, b) a restauração dos dados em ambiente local, c) a adequação do formato do Banco de Dados para manipulação.

Passo 1: Extração da informação do Banco de Dados do Hospital (aquisição)

Os dados para estudo estão armazenados em um sistema de banco de dados de grande porte. Após cuidadosa análise, no sentido de preservar o sigilo pessoal dos pacientes e excluir dados sensíveis, a extração foi realizada através de um filtro do sistema de BI do hospital, já mencionado. Os dados foram selecionados a partir de um recorte conceitual, a saber, “Anamnese”. A equipe de TI do hospital optou por exportar os dados para um banco de dados relacional de menor porte. Este último artefato foi enviado em formato compactado à pesquisadora como fonte de dados principal para a pesquisa.

Passo 2: Restauração dos dados em ambiente local

Após o recebimento do arquivo com os dados em formato compactado, foi realizada instalação de software servidor de banco de dados local (PostgreSQL), com o objetivo de restaurar os dados e possibilitar sua consequente manipulação.

Passo 3: adequação do formato do Banco de Dados para manipulação

A análise preliminar dos dados levou a uma decisão de simplificar o acesso aos mesmos, exportando-os para um banco de dados relacional (SQLite) que não exige um servidor de banco de dados. Isso permitiu que os algoritmos de análise, na linguagem *Python*, tivessem acesso direto e simplificado aos dados. A partir deste formato, foram realizados os processos de pré-tratamento e análise.

Passo 4: pré-processamento

A etapa denominada pré-processamento é merecedora de especial atenção: é a partir da transformação que se processa nessa etapa que os dados são preparados para análise pelos demais algoritmos. A linguagem natural usada na descrição médica proporciona um texto não-padronizado, em forma e em sintaxe, que necessita de intervenção antes da extração em si. Nesta etapa pode-se elencar:

- Retirar quebras de linha: os textos originais são formatados com quebras de linhas para facilitar o entendimento humano, mas tais quebras não são necessárias ao processamento computacional. Os caracteres para esta formatação "/n" são excluídos tornando o texto uma sequência de caracteres, usualmente denominada *string*.

- Retirar caracteres especiais e pontuação: alguns caracteres como traço "-", ponto especial "•", e as sinais definidos pela constante *Python punctuation* – à saber, = " " ! # $ % & ' () * + , - . / : ; <=> ? @ [\] ^ _ `{ | } ~ – são retirados do texto original.

- Retirar excesso de espaços: ao retirar os caracteres indesejados, ou mesmo a digitação original do texto, pôde-se perceber mais de um caractere espaço separando as palavras; uma expressão regular foi usada para normalizar os espaços entre os tokens.

- Transformar todo texto em minúsculas: a fim de padronizar todo o texto foi usado um *case folding* que transforme todos os caracteres do texto para a forma minúscula.

O resultado desta etapa é um novo texto, armazenado em uma nova coluna no banco de dados. Esta coluna será usada na próxima etapa, a qual envolve a extração e a análise de informação.

Passo 5: extração

Na etapa de extração de informação foram desenvolvidos algoritmos na linguagem *Python* para extrair:

- Frequência de CIDs;
- Frequência de *stop-words*;
- Frequência bigramas e trigramas;
- Quantidade de informação afirmativas e negativas.

Para a última tarefa da lista acima foi construída uma lista de termos para delimitar o algoritmo na busca por informação negativa e afirmativa sobre o assunto ginecologia. Foram criadas listas de terminologias junto a equipe do Núcleo Integrado de Pesquisa e Tratamento da Endometriose (NIPTE)[5] do HFR. Tais se baseavam em terminologias utilizadas nos formulários, que o NIPTE utiliza do REDCap[6] para coletar dados de pesquisas científicas sobre cirurgia de Endometriose, Histeroscopia e Miomectomia.

Para as listas de terminologias da temática de obstetrícia foram utilizados os termos de protocolos clínicos e manuais de ginecologia tais como: SES/MG (SES/MG, 2013), Peixoto (2014), Brasil (2016), Brasil (2017), CONITEC (2016), Matos et al. (3017).

Os algoritmos foram desenvolvidos em funções independentes de forma que, a partir dos dados processados, realizam etapas para a conclusão de cada objetivo. No

[5]Disponível em: https://www.feliciorocho.org.br/endometriose
[6] Disponível em : https://redcap.feliciorocho.org.br/redcap/index.php

caso dos três processos que dependem de listas, recuperam-se os dados a partir de arquivos em texto livre. Como padrão de saída para facilitar a análise por pessoas, os dados foram gravados em planilhas eletrônicas.

4 RESULTADOS

Ao extrair dados de anamneses da ginecologia foram recuperados 18.341 documentos. Erros na extração se referiam, principalmente, a digitação, como por exemplo: "qie ": 1, "esv": 1; “secvreçao”:1. A seguir são apresentados os resultados da pesquisa:

a) Presença de CID: foram identificadas notações referentes à CID nos registros de anamnese segundo a Quadro 1:

Quadro 1 - Siglas CID encontradas nos documentos

Código encontrado	Descrição
C19	Neoplasia maligna da junção retossigmóide
E03	Hipotireoidismo congênito com bócio difuso
M88	Doença de Paget do osso (osteíte deformante)
C56	Neoplasia maligna do ovário
C80	Neoplasia maligna, sem especificação de localização
D27	Neoplasia benigna do ovário

b) Frequência de *Stop Word*:

Figura 2 - *Stop Words* da Anamnese

Fonte: Dados da pesquisa, 2020.

c) Frequência de trigramas:

Tabela 1 – Trigramas mais frequentes na anamnese

Tri-gramas	Frequência
em,uso,de	1133
cm,de,vol	1061
ao,exame,mamas	812
normais,abdome,livre	547
hpp,nega,comorbidades	530
mamas,normais,abdome	520
18,utero,de	507
abdome,livre,colo	499
anexos,livres,cd	487
vida,sexual,ativa	480

Fonte: Dados da pesquisa, 2020.

d) Frequência de trigramas de expressões afirmativas e negativas:

Tabela 2 - Expressões afirmativas/ negativas

Expressão afirmativa	Frequência	Expressão negativa	Frequência
normais,abdome,livre	547	hs,nega,tabagismo	399
abdome,livre,colo	499	negativo,para,neoplasia	398
anexos,livres,cd	487	colo,schiller,negativo	387
ca,de,mama	442	livre,colo,schiller	378
avf,tc,normais	437	exame,mamas,sem	377
livres,cd,co	426	toque,sem,alteração	341
mamas,normais,vulva	278	mamas,sem,alterações	316
normais,vulva,ok	268	nega,comorbidades,alergia	297
abdome,livre,vulva	252	sem,alteração,cd	297
mm,ovários,normais	221	negativo,toque,sem	260
exame,mamas,vulva	216	shiler,negativo,toque	217
anos,prevenção,hp	204	sem,sinais,de	214
ultima,consulta,ginecologica	188	colo,shiler,negativo	211
paciente,admitida,para	182	fisiologico,schiller,negativo	178
normal,tq,colo	174	intestinais,preservados,nega	173
do,colo,uterino	173	sem,alterações,abdome	165
fez,uso,de	169	a,palpação,sem	118

exame,beg,hidratada	165	palpação,sem,sinais	118

Fonte: Dados da pesquisa, 2020.

5. DISCUSSÃO

Após as análises iniciais do banco de dados, percebeu-se uma ausência de padrão de nomes para representar os documentos de anamneses (vide TABELA 3). A diversidade de nomes para representar documentos eletrônicos, os quais tem a a mesma finalidade, corrobora com a mencionada complexidade em extrair dados de prontuários. Identificou-se ainda que alguns documentos foram criados, mas não foram preenchidos, demonstrando a necessidade de se realizar curadoria e gestão de documentos em PEPs.

A falta de padronização para geração de documentos eletrônicos permitiu que a equipe médica solicitasse ao setor de TI a criação de documentos no PEP, os quais foram posteriormente pouco utilizados pela dificuldade na identificação e na recuperação dos mesmos. Para a presente pesquisa, optou-se pelos documentos que apresentaram o termo “anamnese” em seu descritivo. A tabela 3 permite visualizar as variações de nomenclatura encontradas no banco de dados do PEP.

Tabela 3 – Frequência de documentos preenchidos pela ginecologia em 2018

DOCUMENTO	FREQUÊNCIA
ANAMNESE - EXAME FISICO	18256
ANAMNESE E EXAME FISICO	10
ANAMNESE GINECOLOGICA	75

Fonte: Dados da pesquisa extraídos do MV-PEP do HFR (2020).

As análises iniciais e a extração dos dados de PEPs foram importantes para identificar os principais problemas de informação do sistema e realizar recuperação de informação de forma assertiva. Os dados foram analisados junto à equipe de ginecologia para fins de validação, correção e melhorias no algoritmo de extração.

O algoritmo e a extração de dados utilizando as técnicas de PLN foram armazenados em repositório digital no GitHub[7]. Para controle de tarefas em equipe foi utilizado o software Trello. Outras ferramentas relevantes no desenvolvimento, algumas das quais já citadas foram:

- PostgreSQL: servidor de banco de dados para restaurar os dados;
- *Visual Studio Code*: ambiente de desenvolvimento para desenvolvimento.
- *Dbeaver*: interface para interação com os bancos de dados;
- Google Drive: software para centralizar arquivos "em nuvem", para fins de compartilhamento de dados e *backup*.

6. CONSIDERAÇÕES FINAIS

A aplicação de técnicas de PLN foi possibilitou delimitar e testar tarefas em anotações em grandes volumes de anamneses de PEP. Por meio dos dados provenientes da anamnese foi possível levantar características dos pacientes para tomada de decisão, por exemplo, a inclusão de pacientes em estudos clínicos na instituição. Além disso, esses dados são fontes para realização de pesquisas clínicas, acadêmicas e estudos epidemiológicos.

A partir da extração dos termos presentes na anamnese, o próximo passo da pesquisa foi verificar correlações com ontologia biomédica. Dessa forma, esperava-se avaliar a real correspondência entre termos do jargão médico presente nas anamneses e os termos canônicos encontrados em terminologias clínicas formais, como as ontologias.

A gestão hospitalar carece em estabelecer diretrizes corporativas que promovam a padronização de processos na criação de documentos e formulários do PEP em relação a anamnese. Como produto da pesquisa espera-se criar um léxico computacional, em português, para delimitar o algoritmo no domínio da ginecologia, extrair relações formais e termos para promover a interoperabilidade entre terminologias e possibilitar o enriquecimento de ontologias biomédicas.

[7] https://github.com/amandadsouza/RiLN

REFERÊNCIAS

Almeida, M. B. **Ontologia em Ciência da Informação:** Teoria e prática. Curitiba, Brasil: CRV, 2020. v. 1, (Representação do Conhecimento em Ciência da Informação).

Almeida, M.B., Souza R.R., Baracho, R.M.A. **Looking for the identity of Information Science in the age of big data, computing clouds and social networks**.Status.In: Proceedings of the 4th International Symposium of Information Science (ISI 2015), Croatia,2015.

Barion, E.C.N., Lago, D. Mineração de textos. **Revista de Ciências Exatas e Tecnologia**, 2008; 3 (3): 123-140.

Baneyx, A., Charlet, J., Jaulent, M.C. **Methodology to build medical ontology from textual resources**. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2006: 21-25.

Blake, C. **Information retrieval**. ARIST, 2011: 45(I sec. II):121-155.

Brasil. Resolução CFM nº 2056 de 20/09/2013. [...]. **D.O.U**, Brasília, 12 nov 2013. [citado 2020 já, 7].Disponível em: https://www.legisweb.com.br/legislacao/?id=261676.

Brasil. **Protocolos da Atenção Básica**: Saúde das Mulheres. Brasília: Ministério da Saúde, 2016.230 p.

Brasil. Secretaria de Atenção à Saúde. **Manual de acolhimento e classificação de risco em obstetrícia**. Brasília: Ministério da Saúde, 2017.64 p.

Campos MLA. **Linguagem documentárias**: teorias que fundamentam sua elaboração. Niterói (RJ): EdUFF; 2001. 133p.

Comissão Nacional de Incorporação de Tecnologias (CONITEC). **Diretrizes Nacionais de Assistência ao Parto Normal**. Brasília: Ministério da Saúde; maio 2016.399p. (Relatório de recomendação, nº 211).

Conselho Federal de Medicina. **Código de Ética Médica**. Resolução CFM nº 1.931, de 17 de setembro de 2009.Brasília: CFM, 2010.[citado 2020 jan 7]. Disponível em: https://portal.cfm.org.br/images/stories/biblioteca/codigo%20de%20etica%20medica.pdf.

Ciol, R., Beraquet, V.S.M. O. Evidência e informação: desafios da medicina para a próxima década. **Perspectivas em Ciência da Informação**, 2009; 14(3):221-230.

Dalianis, H. **Characteristics of Patient Records and Clinical Corpora**. In: Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. 2018b. cap. 4.[citado 2019 jan 2]. Disponível em: <http://link.springer.com/10.1007/978-3-319-78503-5>.

Dias-Da-Silva, B.C. **O estudo linguístico-computacional da linguagem**. Letras de Hoje, 2006: 41(2):103-138.

Farlex Partner Medical Dictionary. **Anamnesis**. 2012.[citado 2020 jan 7]. Disponível em: https://medical-dictionary.thefreedictionary.com/anamnesis.

Freitas, F., Schulz, S., Moraes, E. Pesquisa de terminologias e ontologias atuais em biologia e medicina. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúde**; 3(1):2009.doi:https://doi.org/10.3395/reciis.v3i1.816

Friedman, C., Hripcsak, G. Natural language processing and its future in medicine. **Acad Med**., 1999; 74(8);890-5.

Grüne S. Anamnese und körperliche Untersuchung. **Dtsch Med Wochenschr**., 2016 jan; 141(1):24-7.

IBM. Knowledge Center. **Tokenização**.[citado 2018 nov 8]. Disponível em; https://www.ibm.com/support/knowledgecenter/pt-br/SSPT3X_4.1.0/com.ibm.swg.im.infosphere.biginsights.text.doc/doc/ana_txtan_tokenization.html.

López M. **Anamnese**. In: López M, Medeiros JL. Semiologia Médica: as bases do diagnóstico clínico. 3.ed. Atheneu: Rio de Janeiro; 1990. Cap.2, p.20-34.

Kim Y.S., Yoon D., Byun J., Park H., Lee A., Kim I.H. Extracting information from free-text electronic patient records to identify practice-based evidence of the performance of coronary stents**. PLoS One**. 2017 Aug 11;12(8):e0182889. doi: 10.1371/journal.pone.0182889.

Manning, C.D., Schütze, H. **Foundations of statistical natural language processing**. Massachusetts: MIT press; 1999.620p.

Matos, MS., et al. **Manual de Ginecologia**. Salvador: EBMSP; 2017.

Meystre, S.M., Friedlin, F.J., South, B.R., Shen S, Samore, M.H. Automatic de-identification of textual documents in the electronic health record: a review of recent research. **BMC Med Res Meth.** 2010 Aug 2;10:70. doi: 10.1186/1471-2288-10-70.

Peixoto, S. **Manual de assistência pré-natal.** 2.ed. São Paulo: Federação Brasileira das Associações de Ginecologia e Obstetrícia (FEBRASGO); 2014.

Schulz, S., Rodrigues, J.M., Rector, A., Chute, C.G. Interface Terminologies, Reference Terminologies and Aggregation Terminologies**:** A Strategy for Better Integration. **Stud Health Technol Inform**. 2017;245:940-944.

Spackman, K.A., Hersh, W. R. Recognizing noun phrases in medical discharge summaries: an evaluation of two natural language parsers. **Proc AMIA An. Fall Symp**, 1996:155-158.

Santos, R.E.S., Souza E.P.R., Correia-Neto, J.S., Magalhães, C.V.C., Vilar, G. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. **Revista de Sistemas e Computação**, 2015; 4(2).

SES/MG. Associação de Ginecologistas e Obstetras de Minas Gerais – SOGIMIG. **Atenção à saúde da gestante.** Novos Critérios para Estratificação de Risco e Acompanhamento da Gestante: PROGRAMA VIVA VIDA Projeto Mães de Minas. Maio 2013.

Zhou, X., Han, H., Chankai, I., Prestrud, A.A., Brooks, A.D. **Approaches to Text Mining for Clinical Medical Records**. The 21st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Computer Applications in Health Care (CAHC 2006), Dijon, France, April 23 -27, 2006:235-239. [citado 2019 jun 20]. Disponível em: http://www.ischool.drexel.edu/faculty/hhan/SAC2006_CAHC.pdf.

Wallach, H.M. **Topic Modeling**: Beyond Bag-of-Words. In: Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, 2006.

Wang, Z., Shah, A.D., Tate, A.R., Denaxas, S., Shawe-Taylor, J., Hemingway, H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. **PLoS One**. 2012;7(1):e30412. doi: 10.1371/journal.pone.0030412.