MedicalSearch: Um Sistema Web para Extração de Informação de Artigos da Área Médica

Arthur Emanuel de Oliveira Carosia

Instituto Federal de São Paulo (IFSP), São João da Boa Vista, SP, Brasil

Abstract. PubMed Central is an online digital archive of medical papers, which currently contains more than two hundred thousand papers. It is a valuable source of information in the medical field. However, it does not offer functionalities that allow medical experts to automatically identify and extract key information related to issues of interest from these papers. In this paper, we propose *MedicalSearch*, a web-based system for extracting information from medical papers, which is aimed at aiding medical experts to find information of interest in scientific papers that are stored in PubMed Central. *MedicalSearch* allows medical experts to search for treatments and causes of complications, returning sentences that can be further explored by medical experts according to their needs. In the evaluating process, *MedicalSeach* was tested with real complications of the Sickle Cell Anemia Disease and obtained precision values between 50% and 66%, besides recall values between 63% and 75%.

Resumo. *PubMed Central* é uma base online de artigos da área médica que possui atualmente mais de duzentos mil artigos, se tratando de uma importante fonte de informação dessa área. Entretanto, uma de suas limitações é não oferecer funcionalidades que permitam aos especialistas médicos automaticamente identificar e extrair informações de interesse desses artigos. Este artigo apresenta o sistema *MedicalSearch*, um sistema *web* para extração de informações de artigos da área médica, cujo objetivo é auxiliar especialistas médicos a encontrarem informações de interesse em artigos que estejam armazenados no PubMed Central. MedicalSearch permite aos especialistas da área médica procurarem por tratamentos e causas de complicações ou doenças em artigos da área médica, retornando sentenças para serem exploradas por especialistas da área de acordo com sua necessidade. No processo de avaliação, *MedicalSearch* foi testado levando-se em consideração complicações reais da doença Anemia Falciforme e alcançou valores de precisão que variam de 50% a 66%, além de valores de revocação que variam de 63% a 75%.

***Index Terms*— Medical Information Systems, Computer Applications**

I. Introdução

PubMed Central [17] é um acervo digital de artigos da área médica, que atualmente contém mais de duzentos mil artigos. Atualmente, esta grande quantidade de artigos científicos pode ser pesquisada por meio de funcionalidades simples como, por exemplo, os artigos podem ser recuperados de acordo com os nomes de seus autores, ano de publicação, nome da revista e palavras que estão presentes no título do trabalho. Uma vez encontrado, o conteúdo desses documentos deve ser investigado manualmente pelo especialista para encontrar as informações necessárias. Além disso, deve-se levar em consideração também que a busca manual por informações importantes relacionadas a assuntos de interesse dentro do conteúdo de artigos é passível de erro, especialmente quando o número de artigos é elevado. Portanto, existe a necessidade de desenvolver técnicas e ferramentas que auxiliem os especialistas médicos a identificar automaticamente e extrair informações de artigos da área médica.

Dessa forma, este artigo aborda a proposta de *MedicalSearch*, um sistema *web* para extrair informações de artigos do domínio médico, visando auxiliar especialistas médicos a buscarem informações de interesse em artigos armazenados no *PubMed Central*. Os assuntos de interesse extraídos pelo sistema referem-se a tratamentos e causas de doenças ou complicações. Diferentemente das funcionalidades de pesquisa fornecidos pelo *PubMed Central*, *MedicalSearch* investiga principalmente o conteúdo dos artigos científicos, ou seja, ele analisa todas as sentenças de todos os parágrafos que compõem cada artigo na busca de informações relacionadas a assuntos de interesse. Considerando uma dada complicação de uma doença definida por um especialista da área médica como, por exemplo, hipertensão. O sistema *MedicalSearch* percorre uma base de artigos da área médica verificando se existe uma causa ou tratamento relacionado para esta complicação. Por fim, o sistema retorna ao especialista todas as frases de artigos que contém informações importantes que foram recuperadas nesse processo. O sistema também permite que o especialista da área médica visualize os artigos originais a partir dos quais as informações foram extraídas.

As vantagens práticas do uso de *MedicalSearch* são muitas. É um sistema *web* que fornece um processo ágil de extração de informação, além de ser menos propenso a erros em relação ao processo de extração manual feita por um especialista. O sistema também pode ser designado para uso de estudantes de medicina, que podem usar o sistema para investigar rapidamente o estado da arte de uma determinada doença. Além disso, o sistema *MedicalSearch* pode ser usado por médicos não-especialistas em uma determinada área, tais como clínicos e pediatras gerais, que podem precisar de informações específicas sobre uma determinada doença

Arthur Emanuel de Oliveira Carosia é bacharel e mestre em Ciência da Computação pela Universidade de São Paulo. Atualmente é professor do ensino básico, técnico e tecnológico no Instituto Federal de São Paulo, Câmpus São João da Boa Vista (e-mail: arthuremanuel.carosia@gmail.com).

rapidamente, a fim de tratar um paciente na sala de emergência, por exemplo.

Resumidamente, *MedicalSearch* extrai a informação relacionada a causas e tratamentos de doenças ou complicações a partir de artigos da área médica. Para isso, o sistema possui implementado regras de extração por meio do uso de expressões regulares na identificação dos termos mais relevantes contidos nas sentenças dos artigos, além de uma interface gráfica desenvolvida para o especialista médico.

Este artigo está organizado como descrito a seguir. A Seção 2 resume a revisão bibliográfica sobre extração de informação e os trabalhos relacionados. A Seção 3 descreve o desenvolvimento deste projeto e a Seção 4 apresenta a avaliação da proposta realizada. Finalmente, a Seção 5 conclui este trabalho e apresenta propostas para trabalhos futuros.

II. Revisão Bibliográfica

Esta seção aborda a revisão bibliográfica realizada para o desenvolvimento deste trabalho e está organizada como apresentado a seguir. A Seção 2.1 descreve conceitos relacionados a extração de informação e a Seção 2.2 aborda os trabalhos relacionados.

A. Extração de Informação

A extração de informação tem como objetivo o reconhecimento de trechos de texto em documentos escritos em linguagem natural, assim como também a extração de informação estruturada de documentos não-estruturados [8], [2]. De acordo com [4], o objetivo da pesquisa em extração de informação é desenvolver sistemas que consigam extrair e combinar informações relevantes enquanto ignoram as informações irrelevantes. A informação a ser extraída desses sistemas é definida por regras ou padrões, que dependem do domínio no qual está inserido. Desse modo, estes padrões necessitam ser definidos manualmente por algum especialista.

Na literatura existem três principais abordagens para extração de informação: baseada no uso de dicionários, baseada em aprendizado de máquina e baseada em regras [1]. A informação extraída baseada em dicionários usa uma lista de termos para identificar suas ocorrências em sentenças do texto. Isso é feito a partir da comparação de termos contidos no dicionário com palavras que formam as sentenças. A abordagem baseada em aprendizado de máquina é baseada em princípios de inteligência artificial. Nessa abordagem, um sistema de aprendizado analisa e generaliza a informação, que será utilizada para obter novo conhecimento por meio de inferência lógica. Por fim, a abordagem baseada em regras tem seu princípio na criação de padrões que serão comparados com termos das sentenças dos documentos nos quais a informação de interesse será extraída. Além disso, os padrões podem ser implementados por meio do uso de expressões regulares, como é descrito neste trabalho. O sistema *MedicalSearch* proposto nesse artigo utiliza a abordagem baseada em regras. A motivação para uso dessa abordagem é relacionada ao fato da literatura destacar que esta proposta produz melhores resultados do que as demais abordagens [1], [18].

B. Trabalhos Relacionados

Existem estudos na literatura médica que extraem informações de interesse de artigos da área médica, mas eles diferem deste trabalho tanto em sua finalidade como em sua aplicabilidade.

O trabalho [3] apresenta uma arquitetura de sistema de extração de informação da área médica. Este trabalho apresenta um processo para extrair informações relevantes de artigos médicos envolvendo três etapas: extração, geração e aquisição do conhecimento. No entanto, esse trabalho tem algumas limitações: (i) o trabalho foi aplicado apenas para textos médicos franceses e foi testado apenas para registros de colonoscopia; (ii) o sistema tem dificuldades para analisar frases complexas; (iii) este estudo não se concentra no processo de aquisição de conhecimento.

Outro sistema baseado em regras é a *RLIMS-P* [10], cuja finalidade é extrair informações sobre proteínas de fosforilação de resumos de artigos da área. Nesse trabalho, padrões foram criados depois de examinar diferentes formas usadas para descrever as interações de fosforilação em 300 resumos. Apesar deste trabalho apresentar altos valores de precisão e revocação, sua limitação é extrair apenas informações a partir de resumos e sobre fosforilação de proteínas. Assim, não se considera todo o artigo e o conteúdo para extração é reduzido.

O trabalho [12] aborda a extração de informações a partir de textos que descrevem os relatórios de mamografia e registros hospitalares de pacientes diabéticos. Ele baseia-se na utilização de uma ontologia específica e pré-definida, a qual descreve os domínios mencionados para guiar a extração de informação. Portanto, é um trabalho dependente do domínio ao qual está inserido, que não foi concebido para extrair informação de tratamentos e causas de qualquer doença. Além disso, ele extrai informações somente a partir de registros médicos, que são documentos semi-estruturados. Além disso, o trabalho não se concentra em artigos completos.

Outro trabalho relacionado é descrito em [13]. Seu objetivo é extrair informações de descrições de problemas médicos e tratamentos de dados de fóruns médicos. Este trabalho usa métodos de aprendizado de máquina para realizar suas atividades. Uma limitação deste trabalho é que ele extrai informação somente de fóruns na Internet, que não são fontes de informação tão confiáveis como os artigos da área médica armazenados no *PubMed Central*. Além disso, como este trabalho é baseado em algoritmos de aprendizado de máquina, ele precisa de um conjunto com muitas sentenças para a etapa de treinamento, para que sejam obtidos bons resultados na etapa de extração.

Por outro lado, o trabalho descrito em [11] aborda testes com quatro motores que realizam extração de informações de documentos clínicos. Dois mil resumos e anotações foram submetidas aos quatro motores com o objetivo de extrair todas as informações das medicações utilizadas. Após comparação dos resultados obtidos com a extração, o trabalho concluiu que as ferramentas de extração automatizadas não estão prontas para automatizar a extração de dados de medicamentos sem intervenção humana.

Finalmente, *MedLEE*, [5] e [6], é uma ferramenta especializada no domínio médico que utiliza processamento de linguagem natural. *MedLEE* possui um analisador que detecta

a estrutura gramatical das sentenças, mapeia a estrutura encontrada e as preenche com frases. Como resultado, a saída da ferramenta *MedLEE* contém informações identificadas a partir de prontuários médicos (ex. problemas do paciente) em forma estruturada. No entanto, *MedLEE* tem a limitação de não extrair informações sobre tratamentos para os pacientes e, além disso, é restrito à extração de informações apenas a partir de registros de pacientes não considerando artigos médicos. Por outro lado, *MedicalSearch* extrai informações de questões que estão mais próximos à rotina de médicos especialistas, como causas e tratamentos de doenças ou complicações, através da análise de artigos completos. Além disso, ele pode ser aplicado para encontrar informações em qualquer área do domínio médico, independentemente da doença.

Diferentemente dos trabalhos analisados, *MedicalSearch* realiza a extração de informação relacionada a causas e tratamentos de doenças ou complicações e ainda oferece uma interface *web*, através da qual especialistas médicos podem visualizar as sentenças extraídas e analisar os artigos originais de acordo com as suas necessidades. As inovações introduzidas neste trabalho são: regras criadas para extração de tratamento e causa, o desenvolvimento de um algoritmo de extração, a arquitetura para o sistema *web* e também a metodologia de desenvolvimento.

III. A Ferramenta Medical Search

Nesta seção, é apresentado o desenvolvimento do sistema *MedicalSearch*, um sistema *web* para extração de informação de causas e tratamentos de artigos da área médica. Para isso, o sistema faz uso de regras definidas a partir de expressões regulares relacionadas com a identificação de causas e tratamentos de doenças ou complicações. Assim, o sistema *MedicalSearch* introduz as seguintes características:

- Identificação de padrões relacionados a causas e tratamentos em artigos da área médica de língua inglesa.
- Especificação de regras de extração com o uso de expressões regulares a partir dos padrões identificados.
- Introdução de um algoritmo que percorre os artigos da area médica identificando e extraindo informações de interesse.

Esta seção está organizada como descrito a seguir. A Seção 3.1 descreve a arquitetura do sistema MedicalSearch, a Seção 3.2 define as regras criadas para extração de informação, a Seção 3.3 aborda o algoritmo de extração desenvolvido e a Seção 3.4 descreve a interface com o especialista.

A. Arquitetura

A arquitetura do sistema *MedicalSearch* é composta por três módulos principais:

- Módulo de *download* automático de artigos.
- Módulo de pré-processamento.
- Módulo de extração de causas e tratamentos.

A Figura 1 ilustra a arquitetura do sistema.

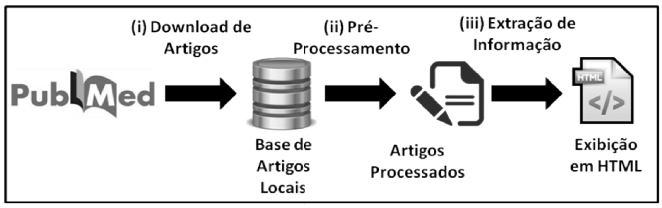

Figura 1. Arquitetura do Sistema

O primeiro módulo (i) é o responsável por realizar automaticamente o *download* de artigos do site *PubMed Central*. *PubMed Central* é um site que contém centenas de milhares de artigos da área médica, sendo que parte destes artigos estão disponíveis para *download* gratuito. Por meio de um cliente FTP ou mesmo de um navegador de internet, é possível ser realizado o *download* destes artigos. No entanto, o processo manual é exaustivo e pode requerer um tempo proibitivo para ser executado. A fim de diminuir esse tempo e poupar esforço dos especialistas, foi desenvolvido um módulo de *download* que realiza *download* automático dos artigos e os armazena em uma base de artigos locais.

O segundo módulo (ii) realiza o pré-processamento dos artigos que estão armazenados localmente a partir do resultado obtido no módulo anterior. Como os artigos estão normalmente armazenados no formato HTML, é necessário que se retire de suas etiquetas (i.e., *tags* HTML*)*, bem como também de códigos de linguagem *CSS* e *Javascript* que estão contidos em seus códigos-fonte e que devem ser eliminados no momento da extração de texto. O principal objetivo do segundo módulo é, portanto, deixar o artigo apenas no formato texto, para que a extração não apresente caracteres ou códigos indesejados.

Finalmente, o terceiro módulo (iii) tem a principal função do sistema elaborado: a extração de informação de artigos da área médica. A extração é feita a partir dos artigos no formato texto obtidos no módulo anterior. Desse modo, o sistema é composto de regras descritas em expressões regulares, descritas na Seção 3.2, que têm como objetivo a extração de causas e tratamentos de doenças ou complicações. Além disso, esse módulo possui uma interface *web* com o especialista, descrito na Seção 3.4, de modo que o especialista possa decidir qual complicação buscar nos artigos. Por fim, o terceiro módulo apresenta o resultado da extração em formato *web*, visando facilitar a visualização das informações extraídas. Ele também oferece a possibilidade do usuário médico retornar ao artigo original, do qual a informação foi retirada.

B. Identificação de Padrões e Regras de Extração

A identificação de padrões de artigos da área médica foi realizada como descrito a seguir. A partir de uma seleção de 50 artigos em inglês da área médica, foram separadas manualmente várias sentenças que continham causa ou tratamento para alguma doença ou complicação em pacientes. Esse processo foi realizado com o auxílio de especialistas da área médica para validação das sentenças. Em seguida, estas sentenças foram classificadas de acordo com algum padrão que pudesse ser utilizado para extraí-las automaticamente de um artigo. Por fim, a partir nos padrões identificados,

realizou-se a construção de regras para extração de causas e tratamentos por meio de expressões regulares.

Padrões para identificar sentenças que contenham tratamentos baseiam-se no fato de que essas sentenças contêm conjugações do verbo tratar, palavra terapia ou variações da palavra *dose*. A partir desse princípio, desenvolveram-se quatro regras definidas com expressões regulares para a extração dessas sentenças. Em todas essas regras, os caracteres .* indicam qualquer conjunto de caracteres, o caractere | indica a operação lógica *ou*, o caractere *?* indica zero ou uma ocorrência do elemento anterior, e os caracteres \w indicam letras ou números. As regras são apresentadas a seguir.

A Regra (1) representa a composição da conjugação do verbo *to treat* (tratar) e a palavra *therapy* (terapia) em conjunto com preposições. Espera-se que o tratamento seja indicado ao final da sentença.

(.*)(treated (by|with) | treatment (with)? | therapy | treating (by|with))(.*) **(1)**

A Regra (2) consiste na conjugação do verbo *to study* (estudar), o termo *trial* (experimento), seguido por uma sequência de caracteres que vai indicar um tratamento ou uma terapia. Espera-se que a descrição do tratamento seja indicada no meio da sentença.

(? : studies|study|studied|studied|trial)(? : [\w − /\\]*)?(? : treatment|therapy) **(2)**

A Regra (3) representa a conjugação do verbo *to receive* (receber) seguido por uma sequência de caracteres e em seguida termos indicando um tratamento, uma terapia ou uma dose de uma droga. Espera-se que a descrição do tratamento seja indicada no meio da sentença.

(? : receive|received|receiving)(? : [\w − /\\]*)?(? : treatment|therapy|dose|doses) **(3)**

A Regra (4) consiste na conjugação parcial do verbo *to treat* (tratar), o termo *threapy* (terapia) e suas preposições. Nesse caso, espera-se que a descrição de o tratamento seja indicada ao final da sentença.

(? : treated|treatment|therapy)(? : with|by)([\w − /\\]*) **(4)**

Em relação à regra de extração de informação de causas de doenças ou complicações, os padrões identificados contêm conjugações do verbo *to cause* (causar) seguido por preposições. Para a extração de frases que contêm causa, foi gerada uma única regra de extração, descrita na Regra (5). Nessa regra, espera-se que a causa de uma determinada doença ou complicação seja indicada ao final da sentença.

(.*)(caused (by|with) | cause (by|with)? | causing (by|with)?)(.*) **(5)**

C. Algoritmo de Extração

O pseudocódigo do algoritmo de extração de informação é demonstrado no Algoritmo 1. Esse algoritmo realiza a extração de informação dos artigos da área médica e utiliza as expressões regulares descritas na Seção 3.2. Seu funcionamento é o seguinte.

O algoritmo recebe como parâmetro uma lista de artigos que terão extraídas suas informações sobre causas e tratamentos de doenças ou complicações, além da complicação que está sendo buscada (linha 1). Em seguida, o algoritmo realiza uma iteração sobre todos os artigos (linha 2) e verifica, para cada sentença do artigo (linha 3), se ela possui a complicação buscada (linha 4) e, em caso afirmativo, se as regras definidas na Seção 3.2 são capazes de extrair alguma informação da sentença atual (linha 5). Se a sentença que está sendo analisada possui informação relevante sobre a complicação, ela é então extraída e inserida a uma lista de sentenças candidatas juntamente com as informações do artigo atual (linha 6). Nessa etapa, são analisadas tanto as regras de extração de tratamentos como a regra de extração de causa. Por fim, o algoritmo retorna uma lista de todos os artigos que possuem sentenças candidatas (linha 9).

```
1 ExtrairInformacao(artigos[],complicação)
2   foreach (artigo in artigos)
3      foreach (sentença in artigo)
4        if (existeComplicação(sentenca)and
5            achouRegra(sentenca))
6            insert(candidatos,sentenca,artigo);
7      end foreach
8    end foreach
9    return candidatos;
10 end
```

Algoritmo 1. Algoritmo de Extração de Informações

D. Interação com o Especialista

Com o objetivo de auxiliar o especialista a visualizar o resultado de uma busca por causas e tratamentos para uma determinada doença ou complicação, o sistema *MedicalSearch* possui também uma interface *web*.

A interface inicialmente aguarda que o especialista digite uma doença ou complicação para a busca. O resultado obtido da extração de causas e tratamentos, por fim, é retornado para análise do especialista. A Figura 2 apresenta a interface inicial da solução desenvolvida, a qual permite ao especialista:

- Digitar em (a) e procurar por causas e tratamentos clicando no botão New Search (b).
- Acessar uma busca previamente armazenada, a qual já foi executada anteriormente, economizando, dessa forma, tempo no processamento ao clicar em Try Previous (c).
- Verificar a palavra na medida em que digita o nome de uma complicação. Isso é feito a partir de uma lista de complicações já armazenadas no sistema e na interface é demonstrado pela caixa com legenda *Suggestions Box* (d).

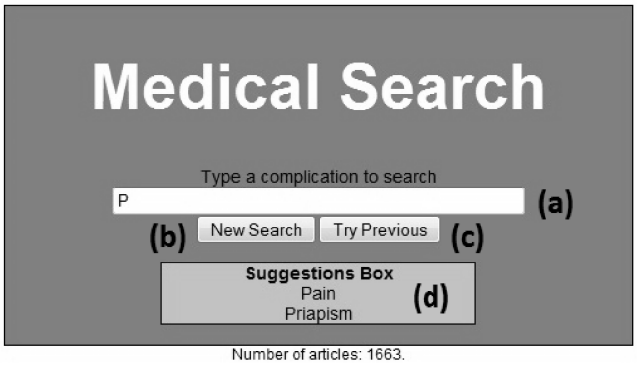

Figura 2. Interface com o Especialista

Já a Figura 3 apresenta o resultado de uma busca feita pelo especialista, a qual exibe causas e tratamentos para a complicação procurada para cada um dos artigos presentes do site *PubMed Cental* que foram abaixados para a ferramenta. Para cada artigo que possui informação relevante (a), as sentenças que possuem indicativo de tratamento são indicadas por *Treatment* e as sentenças que possuem indicativo para causa são indicadas por *Cause*. Para cada sentença, são exibidas suas informações (b) e a o conteúdo extraído (c). Além disso, também existe a funcionalidade que permite que o especialista acesse o artigo original de onde foram extraídas as informações exibidas na interface.

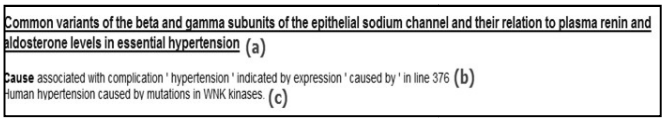

Figura 3. Resultado de Resultado de Busca feita pelo Especialista

Além da interface *web* para o especialista, foi também elaborado um módulo para o que usuário administrador possa baixar artigos automaticamente do *PubMed Central* e também processá-los, fazendo com que fiquem no formato adequado para a extração de informação. Desse modo, o sistema tem dois tipos de usuário, cada qual desempenhando uma função específica. Enquanto o administrador pode tanto buscar como realizar o *download* de artigos e processá-los, o usuário especialista pode apenas realizar a busca por complicações no sistema.

IV. AVALIAÇÃO DA PROPOSTA

Para avaliação do sistema foi realizado um estudo de caso real considerando a doença Anemia Falciforme, que é uma doença genética, hereditária e hematológica que causa a má formação de glóbulos vermelhos [7], que ilustrada na Figura 4. No Brasil, existem poucas pesquisas no sentido de se encontrar uma solução atenuadora da doença. Esta doença não possui cura atualmente e, portanto, há a necessidade de encontrar meios que possam atenuá-la. Existem vários estudos sobre esta doença que podem ser encontrados em artigos científicos e que relatam descobertas interessantes sobre grupos distintos de pacientes, localizações geográficas e tratamentos [9], [14] e [15]. Os resultados descritos nestes artigos são importantes fontes de informação, que podem ser exploradas por especialistas para auxiliar na identificação das causas da doença, na determinação de tratamentos apropriados, e na produção e utilização de novas drogas.

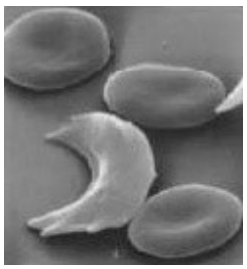

Figura 4. Exemplo de hemácias falciformes [19]

Para realização do processo de avaliação, foi necessário o auxílio de um especialista da área médica. Esse processo foi feito em quatro etapas, descritas a seguir:

1. Seleção aleatória de 15 artigos para composição de um domínio de artigos a serem analisados. Os artigos pertenciam à doença Anemia Falciforme, pois se tratava de uma doença de interesse ao especialista.
2. Extração manual de todas as suas sentenças relacionadas a tratamentos e causas para as complicações priapismo, e hipertensão, que são complicações muito comuns a pacientes da doença Anemia Falciforme.
3. Execução do sistema com apenas esse grupo de artigos;
4. Comparação dos resultados obtidos na etapa de extração manual realizada pelo especialista com os resultados obtidos da execução do sistema.

Para avaliação dos resultados obtidos, as métricas revocação e precisão foram utilizadas. Enquanto revocação é definida pela relação entre documentos relevantes recuperados e o número total de documentos relevantes, precisão é definida pela relação entre documentos relevantes retornados e o número total de documentos retornados [16].

No processo de extração manual com auxílio do especialista, ilustrado na Tabela 1, foram encontradas 24 sentenças relacionadas a tratamentos e 16 sentenças relacionadas a causas para uma complicação típica da doença Anemia Falciforme: priapismo. Com relação ao resultado obtido a partir da execução do sistema, houve o retorno correto de 16 sentenças para tratamento e 12 sentenças para causa. Além disso, 8 sentenças foram retornadas erroneamente para tratamento e 6 sentenças retornadas erroneamente para causas, representando falsos positivos na extração. Desse modo, o sistema apresentou o seguinte resultado para a complicação priapismo. Para tratamento, o valor obtido de revocação foi 66% e o de precisão foi 66%; para causa, o valor obtido de revocação foi 75% e o de precisão foi 66%. Esses resultados são apresentados na Tabela 2.

Por outro lado, para a complicação hipertensão, durante o processo de extração manual, ilustrado na Tabela 3, foram encontradas 12 sentenças relacionadas a tratamentos e 9 sentenças relacionadas a causas. Com relação ao resultado obtido a partir da execução do sistema, houve o retorno correto de 7 sentenças para tratamento e 6 sentenças para

causa. Ainda, 4 sentenças foram retornadas erroneamente para tratamento e 3 sentenças retornadas erroneamente para causas. Assim, os resultados do sistema são apresentados a seguir. Para tratamento, o valor obtido de revocação foi 63% e o de precisão foi 58%; para causa, o valor obtido de revocação foi 66% e o de precisão foi 50%. Os resultados são apresentados na Tabela 4.

Assim, pode-se considerar que essa porcentagem obtida consegue atender as necessidades dos médicos na extração de causas e tratamentos. Além disso, como o sistema também oferece interface ao especialista, ele poderá avaliar qual sentença selecionada poderá ser de seu interesse e ainda assim descartar as sentenças que indicam falsos positivos.

Tabela 1. Retorno de Sentenças para Priapismo

	Total de Sentenças	Sentenças Retornadas Corretamente	Sentenças Retornadas Incorretamente
Tratamento	24	16	8
Causa	16	12	6

Tabela 2. Resultados para Priapismo

	Revocação	Precisão
Tratamento	66%	66%
Causa	75%	66%

Tabela 3. Retorno de Sentenças para Hipertensão

	Total de Sentenças	Sentenças Retornadas Corretamente	Sentenças Retornadas Incorretamente
Tratamento	12	7	4
Causa	9	6	3

Tabela 4. Resultados para Hipertensão

	Revocação	Precisão
Tratamento	63%	58%
Causa	66%	50%

V. Considerações Finais

Neste trabalho, foi apresentado o sistema *web MedicalSearch*, que extrai informações sobre causas e tratamentos de complicações ou doenças de artigos da área médica. Seu objetivo é dar suporte a especialistas médicos a encontrar informações relacionadas com assuntos de interesse em trabalhos científicos armazenados no *PubMed Central*. As principais características do *MedicalSearch* são:

- Identificar padrões e definir regras para extração de sentenças relacionadas a causas e tratamentos de doenças ou complicações em papéis biomédicas.
- Introduzir um algoritmo que percorre artigos da área médica, utilizando as regras de extração para obter as informações de interesse.
- Oferecer uma interface *web* através do qual especialistas da área médica podem procurar por causas e tratamentos em artigos da área.
- Visualizar por meio de interface *web* os artigos originais a partir do qual as informações foram extraídas.

O objetivo da ferramenta *MedicalSearch* é oferecer um sistema *web* que fornece um processo ágil de extração de informação, além de ser menos propenso a erros em relação ao processo de extração manual feita por um especialista. Neste contexto, a principal contribuição do nosso trabalho é, dado o domínio médico, ajudar especialistas a obter resultados relacionados a causas e tratamentos de doenças ou complicações que já foram relatados na literatura. Por exemplo, os resultados fornecidos pelo sistema podem ser usados por especialistas para desenvolver melhores tratamentos utilizando como base existente tratamentos bem sucedidos ou para produzir novos medicamentos de acordo com resultados importantes obtidos a partir de artigos.

Além disso, vale a pena destacar que o sistema *MedicalSearch* foi avaliado levando-se em consideração complicações da doença Anemia Falciforme, apresentando resultados satisfatórios. Os valores de revocação variaram de 63% a 75% e os valores de precisão variaram de 50% a 66%.

Como trabalhos futuros, estão as seguintes tarefas.

- Explorar melhorias do algoritmo de extração de informação dos artigos.
- Validar o sistema com uma quantidade elevada de artigos.
- Refinamento das regras de extração.

Referências

[1] Ananiadou, S.; Mcnaught, J. (Ed.). Text mining for biology and biomedicine. Norwood, MA: Artech House, 2006. 302 p.

[2] D. E. Appelt. Introduction to information extraction. AI Commun. 12(3):161– 1999.

[3] D. Bekhouche, Y. Pollet, B. Grilheres and X. Denis. Architecture of a medical information extraction system. In F. Meziane and E. Métais editors Natural Language Processing and Information Systems. Springer Berlin/Heidelberg, 2004.

[4] Cowie. J., W. Lehnert. Information extraction. Communications of the ACM 39(1), 80-91, 1996.

[5] C. Friedman P. O. Alderson J. H. M. Austin J. J. Ciminoand S. B. Johnson. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1994.

[6] C. Friedman G. Hripcsak L. Shagina and H. Liu. Representing information in patient reports using natural language processing and the extensible markup language. Journal of the American Medical Informatics Association, 1999.

[7] Gladwin, M.T., Kato, G.J. Cardiopulmonary Complications of Sickle Cell Disease: Role of Nitric Oxide and Hemolytic Anemia. Hematology, p. 51-57, 2005.

[8] R. Grishman. Information extraction: techniques and challenges. In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology pages 10–27. Springer Berlin / Heidelberg 1997.

[9] Gulbis, B., et al. Hydroxyurea for Sickle Cell Disease in Children and for Prevention of Cerebrovascular Events: the

BELGIAN EXPERIENCE. CLINICAL OBSERVATIONS, INTERVENTIONS, AND THERAPEUTIC TRIALS, V. 105, N. 7, P. 2685-2690, 2005.

[10] Z. Z. HU M. NARAYANASWAMY K. E. RAVIKUMAR K. VIJAY-SHANKER AND C. H. WU. LITERATURE MINING AND DATABASE ANNOTATION OF PROTEIN PHOSPHORYLATION USING A RULE-BASED SYSTEM. BIOINFORMATICS 21(11):2759– JUNE 2005.

[11] V. JAGANNATHAN C. J. MULLETT J. G. ARBOGAST K. A. HALBRITTER D. YELLAPRAGADA S. REGULAPATI AND P. BANDARU. ASSESSMENT OF COMMERCIAL NLP ENGINES FOR MEDICATION INFORMATION EXTRACTION FROM DICTATED CLINICAL NOTES. I. J. MEDICAL INFORMATICS 78(4):284–2009.

[12] A. MYKOWIECKA AND M. MARCINIAK. DOMAIN MODEL FOR MEDICAL INFORMATION EXTRACTION - THE LIGHTMEDONT ONTOLOGY. IN ASPECTS OF NATURAL LANGUAGE PROCESSING PAGES 333– 2009.

[13] P. SONDHI M. GUPTA C. ZHAI AND J. HOCKENMAIER. SHALLOW INFORMATION EXTRACTION FROM MEDICAL FORUM DATA. IN COLING (POSTERS) PAGES 1158– 2010.

[14] R. E. WARE. HOW I USE HYDROXYUREA TO TREAT YOUNG PATIENTS WITH SICKLE CELL ANEMIA. BLOOD 115(26):5300– 2010.

[15] K.J.WIERENGA I.R.HAMBLETON N.A.LEWIS AND S.C.UNIT. SURVIVAL ESTIMATES FOR PATIENTS WITH HOMOZYGOUS SICKLE-CELL DISEASE IN JAMAICA: A CLINIC-BASED POPULATION STUDY. THE LANCET 357(9257):680–2001.

[16] BAEZA-YATES, RICARDO, AND BERTHIER RIBEIRO-NETO. MODERN INFORMATION RETRIEVAL. VOL. 463. NEW YORK: ACM PRESS, 1999.

[17] PUBMED CENTRAL. DISPONÍVEL EM: HTTP://WWW.NCBI.NLM.NIH.GOV/PMC/. ÚLTIMA VISITA: 29 DE MARÇO DE 2016.

[18] MATOS, P. F. METODOLOGIA DE PRÉ-PROCESSAMENTO PARA EXTRAÇÃO DE INFORMAÇÃO DE DOENÇAS DE ARTIGOS CIENTÍFICOS DO DOMÍNIO BIOMÉDICO. DISSERTAÇÃO DE MESTRADO. UFSCAR.

[19] Rodrigues, C. M. Hemopatias na gestação. 2008. Disponível em: <http://www.paulomargotto.com.br/documentos/Hemopatias%20na%20Gesta%C3%A7%C3%A3o.ppt>. Acesso em: 03 out. 2015.