Sistema de Informação para Perguntas e Respostas em Doenças Crônicas

Luciana F. Almansa[1], **Alessandra A. Macedo**[1]

[1]Departamento de Computação e Matemática – Universidade de São Paulo (USP)
Caixa Postal 14.027-175 – Ribeirão Preto – SP – Brazil

{luciana.f.almansa,ale.alaniz}@usp.br

Abstract. *The search for relevant information considering medical scientific papers is a complex task due to the lack of time and the complexity of creation of queries. Therefore, this work presents a Question Answering (QA) architecture that helps healthcare professionals find related answers quickly. This framework is supported mainly by Text Mining and Information Retrieval techniques and the evaluation is using a chronic disease reference collection and performance measures. This work intents to contribute with a general QA framework architecture to be used in different medical fields.*

Resumo. *No ambiente médico, buscar informações relevantes em artigos científicos é uma tarefa que exige tempo e experiência dos profissionais. Assim, o objetivo deste trabalho é apresentar uma arquitetura de um sistema do tipo Pergunta e Respostas (PR) que auxilie profissionais da área da saúde na busca rápida por respostas. O sistema foi desenvolvido principalmente por técnicas de Mineração de Texto e Recuperação de Informação. A avaliação do sistema está sendo realizada por meio de uma coleção de referência do domínio de doenças crônicas e epigenética e com o uso de medidas de avaliação de desempenho. Este trabalho pretende contribuir com uma arquitetura genérica de sistemas de PR que pode ser adaptada por vários domínios de informação médica.*

1. Introdução

No ambiente médico e de saúde, especificamente no tratamento clínico do paciente, o papel da informação descrita nos prontuários médicos é registrar o estado de saúde do paciente e auxiliar os profissionais diretamente ligados ao tratamento [Bhat et al. 2016]. Os pesquisadores *Shortliffe e Cimino* vislumbram o futuro de sistemas computacionais médicos como *"Learning Healthcare Systems"* (LHS). Segundo esses pesquisadores, sistemas LHS funcionarão de acordo com um ciclo de informação cujos dados distribuídos de prontuários eletrônicos do paciente serão submetidos a base de dados de registros médicos e de pesquisa científica para serem processados e gerarem conhecimento [Shortliffe and Cimino 2013].

Os artigos científicos são uma importante fonte de informações e de descobertas científicas relevantes na área da saúde. Geralmente, os artigos científicos são armazenados em bibliotecas digitais como, por exemplo, o PubMed[1]. À medida que mais artigos

[1]O Pubmed é uma repositório de informações online com mais de 25 milhões de citações, periódicos e livros da literatura biomédica (http://www.ncbi.nlm.nih.gov/pubmed).

são inseridos, a rede de informação vai aumentando e se tornando cada vez mais rica e complexa para a busca da informação desejada [Rzhetsky et al. 2009]. O crescimento acelerado do número de publicações anuais e a falta de tempo dos profissionais da saúde faz com que as informações dos artigos científicos sejam pouco aproveitadas pela sociedade, em geral, e pela comunidade científica e profissional [Shortliffe and Cimino 2013]. Ler artigos científicos é uma tarefa que exige tempo e disposição dos pesquisadores. A busca por informações desejadas tende a ser complexa e minuciosa devido à quantidade de artigos escritos em idiomas diferente da língua nativa e em diferente estruturações e formatos. Neste contexto, os sistemas computacionais de Recuperação de Informação (RI) e de Perguntas e Respostas (PR) foram desenvolvidos.

Os sistemas de RI são responsáveis pela representação e organização das informações de modo que o acesso à informação seja facilitado para o usuário [Baeza-Yates and Ribeiro-Neto 1999]. Uma das desvantagens dos sistemas de RI é a quantidade de documentos retornados como resultado do processo de recuperação de informação [Kolomiyets and Moens 2011]. Em sistemas de PR, a informação que o usuário necessita é apresentada em linguagem natural e, como resultado do processo, respostas curtas e diretas são retornadas ao usuário [Kolomiyets and Moens 2011].

O objetivo deste trabalho é apresentar a arquitetura de desenvolvimento de um sistema de informação do tipo Perguntas e Respostas, o qual coleta perguntas, processa e apresenta respostas a partir da análise de artigos científicos. A arquitetura denominada QASF (*Question Answering Surveillance Framework*) pode ser utilizada em diferentes domínios de informação médica, considerando que os artefatos linguísticos que apoiam o framework como, por exemplo, as ontologias sejam alterados, conforme o domínio de informação escolhido. Inicialmente, o sistema está sendo validado no domínio de doenças crônicas relacionadas a fatores epigenéticos [2]. Este trabalho pretende auxiliar profissionais da área da saúde na busca rápida e direta por informações de interesse.

Este artigo está estruturado na seguinte forma: a Seção 2 apresentada a arquitetura genérica de um sistema de PR e a arquitetura do QASF com a descrição das técnicas utilizadas na construção de cada módulo. A Seção 3 apresentada a avaliação do QASF. Na Seção 4 são detalhados os trabalhos que se relacionam à proposta do QASF e, por fim, a Seção 5 relata as conclusões do trabalho desenvolvido e trabalhos futuros.

2. Arquitetura do Question Answering Surveillance Framework

Os sistemas de informação do tipo PR tem como objetivo fornecer informações diretas e precisas sobre uma pergunta proposta pelo usuário. Para realizar esta tarefa, estes sistemas utilizam técnicas computacionais, principalmente, das áreas de Extração de Informação (EI), Mineração de Texto (MT) e Recuperação de Informação [Allam and Haggag 2012].

O QASF é uma arquitetura do tipo PR que tem como objetivo auxiliar profissionais da área da saúde respondendo questões escritas em linguagem natural realizadas por profissionais dessa área em linguagem natural. O QASF é composto pelos três módulos tradicionais de PR: (i) processamento da questão, (ii) processamento da resposta e (iii)

[2]Estudos em medicina genômica sugerem que seres humanos expostos a fatores de risco no início da vida como, por exemplo, a escassez de alimentos, podem sofrer influências na expressão do gene e, como resultado, na vida adulta, desenvolverem doenças crônicas [Barker 2001]. A área da biologia que estuda estas influências na expressão do gene é conhecida como epigenética [Miyake et al. 2012].

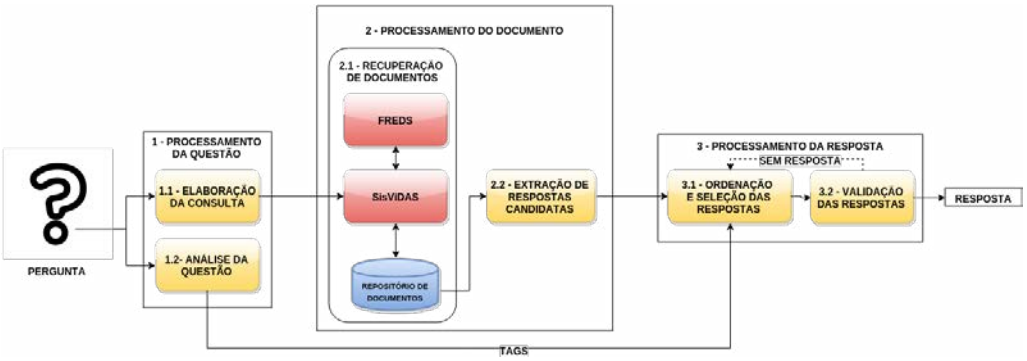

Figura 1. Arquitetura geral de um sistema de PR junto aos submódulos propostos especificamente para o QASF.

processamento do documento; esse último módulo agrega um sistema de suporte que será descrito na Subseção 2.2. A Figura 1 apresenta a arquitetura do QASF que será detalhada nas próximas subseções.

Inicialmente, o domínio de informações em epigenética e doenças crônicas foi escolhido para ser utilizado no desenvolvimento e avaliação deste framework. Contudo, a arquitetura do QASF foi projetada, em sua grande maioria, para ser genérica e suportar diferentes domínios de informações médicas. Porém, o módulo de "Processamento de Documentos" necessita de alguns artefatos linguísticos como, neste caso, uma ontologia do domínio de epigenética e doenças crônicas para realizar uma busca mais apurada pelos artigos que possam conter as respostas corretas. Portanto, caso estes artefatos sejam substituidos, o QASF se torna capaz de responder perguntas em outro domínio médico. Todos os módulos do QASF foram desenvolvidos utilizando a linguagem de computação Python[3], apoiada, principalmente, pelas bibliotecas de Processamento de Linguagem Natural (PLN) *NLTK*[4] e de Aprendizado de Máquina (AM) *scikit-learn*[5].

2.1. Processamento da Questão

O módulo de "Processamento da Questão" objetiva extrair informações da questão realizada pelo usuário em linguagem natural e utilizar as informações na seleção das respostas candidatas. A etapa de "Processamento da Questão" é dividida em dois submódulos: "Análise da Questão" e "Elaboração da Consulta".

A "Análise da Questão" busca informações que auxiliam na localização e na verificação das respostas candidatas. Por exemplo, a pergunta "*O que é doenças crônicas?*" pode extrair informações, por exemplo, o assunto da questão (doenças crônicas). Este tipo de informação (doenças crônicas) pode excluir respostas candidatas cujo assunto seja "câncer". Na literatura, foram encontradas algumas abordagens para "Análise da Questão": correspondência de padrões [Er and Cicekli 2013], Máquina de Vetores de Suporte (do inglês, *Support Vector Machine* - SVM) e análise morfossintática [Monz 2003].

O submódulo de " Análise da Questão" do QASF é suportado por abordagens de

[3]https://wiki.python.org/moin/
[4]http://www.nltk.org/
[5]http://scikit-learn.org/stable/

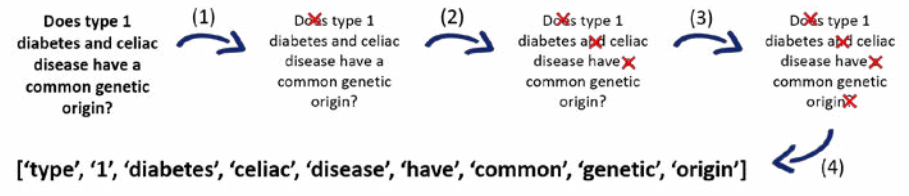

Figura 2. Representação do processo de construção da consulta computacional baseada na pergunta do usuário.

Aprendizado de Máquina, especificamente por SVM e Naive Bayes, para construção de classificadores que categorizam as questões de acordo com o seu assunto. Os classificadores foram construídos e validados utilizando um conjunto de perguntas subdividido de acordo com o assunto da pergunta. Os resultados obtidos foram apresentados e comparados entre si na Seção 3.

A "Elaboração da Consulta" extraí um conjunto de palavras-chave da pergunta feita pelo usuário que será transmitido para o módulo de "Processamento do Documento", especificamente na etapa de "Recuperação de Documentos". No QASF, a construção do vetor de palavras é feita desconsiderando os pronomes interrogativos, as *"stopwords"* e as pontuações. Na Figura 2, o modelo de construção da consulta computacional do QASF a partir de uma pergunta inserida pelo usuário é apresentado. Para iniciar o processo, o usuário insere uma pergunta e como resultado final é construído um vetor com as palavras-chave da pergunta. Na primeira etapa, os pronomes interrogativos são desconsiderados da frase inserida pelo usuário. Em seguida, as *"stopwords"* e, por último, as pontuações da frase são eliminadas. O resultado final é um vetor de palavras-chave que representa a pergunta do usuário.

Os resultados alcançados pelos submódulos "Análise da Questão" e "Elaboração da Consulta" não se relacionam nesta etapa, contudo, eles auxiliam o submódulo de "Extração das Respostas Candidatas" na busca pelas respostas candidatas.

2.2. Processamento dos Documentos

O módulo de "Processamento dos Documentos" engloba os submódulos de "Recuperação dos Documentos" candidatos e de "Extração das Respostas Candidatas". A tarefa de "Recuperação de Documentos" objetiva a recuperação de documentos que contenham possíveis respostas para a pergunta elaborada pelo usuário. Como entrada de dados desta etapa, o vetor de palavras-chave da etapa de "Processamento da Questão" é inserido.

No QASF, o submódulo de "Recuperação de Documentos" foi projetado a partir do SisViDAS [Pollettini et al. 2014]. O SisViDAS é um sistema que relaciona artigos científicos a fatores de riscos descritos em prontuários médico do paciente e alerta profissionais da área da saúde sobre problemas no desenvolvimento humano. Este submódulo foi construído para ampliar o leque das informações manipuladas para a busca de informações relevantes no contexto de um sistema de PR para obter mais precisão das respostas retornadas ao usuário. O SisViDAS compreende os processos de busca da informação em artigos científicos e prontuários médico do paciente nos domínios de epigenética e doenças crônicas.

O submódulo de "Extração das Respostas Candidatas" é um dos mais complexos,

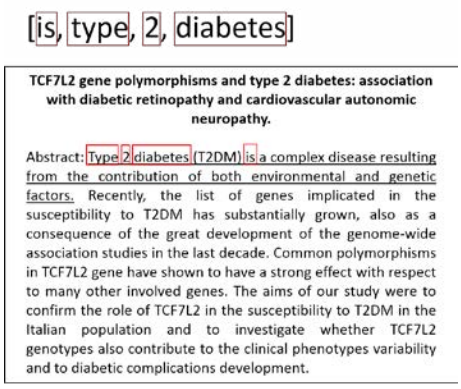

Figura 3. Esquema que mostra a busca por parágrafos candidatos.

pois o usuário pode fazer perguntas em diferentes níveis de complexidade. Na literatura, vários autores propõem diferentes abordagens como, por exemplo, utilizar as palavras encontradas no submódulo de "Análise da Questão", o uso de árvores de análise sintática ou de grafos de dependência [Monz 2003] e a proximidade linear [Gupta and Gupta 2012].

No QASF, a extração de respostas candidatas utiliza os documentos retornados pelo submódulo de "Processamento do Documento". Inicialmente, o conteúdo de cada artigo retornado é segmentado em parágrafos. Em seguida, uma comparação entre o vetor de palavras que representa a pergunta inserida pelo usuário e cada parágrafo segmentado dos artigos retornados pelo submódulo de "Recuperação de Documentos" é realizada. Neste sentido, as palavras contidas no vetor de palavras-chave e no parágrafo candidato ao mesmo tempo são contadas. Caso a quantidade de palavras seja igual ou maior a um liminar l, que pode ser definido manualmente, o parágrafo é considerado um parágrafo candidato. A Figura 3 apresenta um exemplo de resumo de artigo cuja frase sublinhada ilustra o processo de seleção desse parágrafo como resposta candidata.

2.3. Processamento da Resposta

O módulo de "Processamento da Resposta" é a última etapa do processo de PR. Neste módulo, as respostas candidatas são selecionadas e ordenadas de acordo com o grau de similaridade entre a questão e as respostas candidatas. A etapa se divide entre os submódulos de "Seleção e Ordenação das Respostas" e "Validação das Respostas". Como resultado deste módulo, as respostas para a pergunta inserida pelo usuário no início do processo de PR são retornadas. Os sistemas de PR não precisam, necessariamente, retornar uma única resposta, eles podem retornar várias respostas curtas ordenadas de acordo com o grau de relevância. Atualmente, o QASF tem retornado as cinco respostas mais semelhantes à pergunta inserida pelo usuário, contudo, esse valor pode ser reajustado.

O submódulo de "Ordenação e Seleção das Respostas" realiza a ordenação das respostas candidatas, por meio de medidas de similaridade, e apresenta ao usuário um número n de respostas mais semelhantes. Na ordenação das respostas candidatas, alguns critérios podem ser adotados: (i) a quantidade de palavras da questão do usuário que é reconhecida nas respostas candidatas na mesma ordem de posição, (ii) o número de palavras que separam as palavras-chave mais distantes entre si no parágrafo analisado ou (iii) o número de palavras incompatíveis com as palavras-chave [Allam and Haggag 2012].

No QASF, o submódulo de "Ordenação e Seleção das Respostas" realiza o cálculo da similaridade do cosseno. Nesse contexto, a pergunta realizada pelo usuário e as respostas candidatas são convertidas para o espaço vetorial. Desse modo, a similaridade representada pelo cosseno entre a pergunta do usuário e cada resposta candidata é calculada. Desse modo, as respostas são ordenadas e apresentadas ao usuário. A Figura 4 apresenta um exemplo da ordenação realizada pelo QASF.

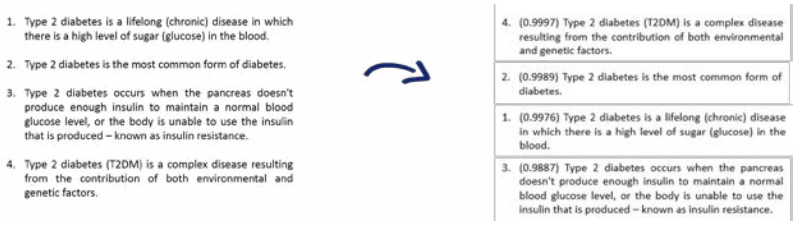

Figura 4. Esquema de ordenação e seleção das respostas candidatas.

3. Avaliação do QASF: Módulo de Análise da Questão

Uma das formas mais conhecidas de avaliação e validação das respostas retornadas pelos sistemas de PR é realizada com o auxílio de *corpora* compostos por tuplas de perguntas e respostas. Os corpora podem ser criados especificamente para a avaliação do sistema em questão ou disponibilizados por conferências como a TREC [Bilotti and Nyberg 2006]. Outras formas de avaliar os sistemas de PR são pela comparação dos resultados de avaliações efetuadas pelo sistema e por um colaborador e pelo uso de FAQs (do inglês, *Frequently Asked Questions*). No contexto das perguntas FAQs, as respostas das perguntas são inseridas em partes aleatórias do documento e o sistema deve retornar a resposta esperada [Lin and Demner-Fushman 2005].

Após a construção dos módulos de processamento da questão, dos documentos e das respostas, a etapa seguinte foi iniciada com as primeiras avaliações do sistema. O primeiro módulo a ser avaliado foi o módulo de análise da questão. Este módulo foi construído com técnicas de SVM e Naive Bayes. *Cao et al.* apresentam uma coleção de referência em seus trabalhos que foi utilizada para a construção e avaliações dos classificadores do QASF [Cao et al. 2011]. Na Subseção 3.1, os detalhes desta coleção de referência são apresentados e na Subseção 3.2, as avaliações e os resultados obtidos são resumidos.

3.1. Coleção de Referência

A coleção de referência usada para avaliar o módulo de "Análise da Questão" do QASF foi construída para avaliação do sistema AskHERMES [Cao et al. 2011] e disponibilizada pelos autores para *download* na web. A coleção de perguntas é composta por 4654 questões médicas sobre doenças crônicas. As questões são separadas em 12 tópicos: dispositivo, diagnóstico, epidemiologia, etiologia, história, gestão, farmacologia, achado físico, procedimento, prognóstico, teste e tratamento e prevenção. As questões foram classificadas pelos próprios clínicos do projeto AskHERMES que fizeram as perguntas. Os tópicos não possuem a mesma quantidade de perguntas. Alguns exemplos de perguntas são: "*What is the protein you can test to see if someone has type 1 or type 2?*". Esta pergunta é do tipo teste. Outro exemplo é "*Is diabetes a risk factor for carpal tunnel syndrome?*" do tipo diagnóstico.

Tabela 1. Classificação de uma coleção de referência em classes pelo QASF.

	Corpus Completo		4258 perguntas – 7 classes	
	Naive Bayes	SVM	Naive Bayes	SVM
Precisão	0.48	0.51	0.48	0.59
Revocação	0.42	0.53	0.46	0.58
F-Measure	0.32	0.51	0.38	0.56
Acurácia	0.43 +- 0.04	0.55+-0.04	0.47 +-0.02	0.59 +-0.03

3.2. Resultados

Os testes realizados com os classificadores e a coleção de referência foram separados em dois cenários para avaliação. No primeiro cenário, toda a coleção de referência é utilizada para avaliar os classificadores Naive Bayes e SVM. No segundo cenário, os tópicos que continham menos de cem perguntas foram desconsiderados. Consequentemente a coleção de referência foi reduzida a 4258 questões distribuídas nos seguintes sete tópicos: diagnóstico (475 perguntas), teste (578 perguntas), tratamento e prevenção (319 perguntas), achados patológicos (189 perguntas), etiologia (172 perguntas), gestão (1.380 perguntas) e farmacologia (1.145 perguntas).

Na Tabela 1, os resultados dos experimentos são apresentados. As duas primeiras colunas compõem o primeiro cenário com toda a coleção de referência para classificação. As duas últimas colunas apresentam o segundo cenário cujos tópicos com menos de cem perguntas são desconsiderados. A tarefa de classificação é avaliada a partir da execução de dois algoritmos de aprendizado de máquina, SVM e Naive Bayes. Em ambos os cenários, o QASF com o classificador SVM apresentou melhor desempenho com valores acima de 0.50 para as medidas de Precisão, Revocação, F-Measure e Acurácia. Especificamente, no segundo cenário, os valores da precisão e acurácia alcançaram valores próximos a 0.6.

Como já apontado em estudos anteriores [Cao et al. 2011, Yen et al. 2013], o classificador SVM apresentou melhor desempenho em ambos os cenários. Além disso, é possível perceber a importância do balanceamento das classes na tarefa de classificação da questão. Esta avaliação indicou performance moderada do QASF. Contudo, um ponto a ser destacado é que as perguntas foram classificadas em seu formato original sem a aplicação de etapas de tratamento da questão como, por exemplo, a eliminação das *stopwords* ou a adição de termos similares por meio da UMLS. Este desempenho moderado pode sobrecarregar ou até atrapalhar os demais módulos do sistema, uma vez que as informações deste módulo auxiliam na escolha das respostas candidatas. Atualmente, os demais módulos estão em avaliação.

4. Trabalhos Relacionados

Na literatura científica atual, diferentes tentativas de construção de sistemas de PR foram identificadas. O primeiro trabalho relacionado apresenta um sistema de PR baseado em abordagens de Aprendizado de Máquina para integração das etapas de classificação da questão e da seleção das respostas. O classificador categoriza a questão inserida pelo usuário e repassa esta informação para a etapa de ordenação das passagens que reorganiza as passagens recuperadas da etapa de processamento dos documentos [Yen et al. 2013]. No domínio médico há dois sistemas que se destacam. O primeiro é o AskHERMES que processa semanticamente questões e busca as respostas em

Tabela 2. Organização das arquiteturas encontradas.

Sistemas de PR	Domínio Informação	Artefatos Linguísticos	Método Classificação
QASF	saúde	sim	SVM e Naive Bayes
[Yen et al. 2013]	genérico	não	SVM
AskHERMES [Cao et al. 2011]	saúde	sim	SVM
MEANS [Ben Abacha and Zweigenbaum 2015]	saúde	sim	NER[6]
[Amorim et al. 2012]	genérico	sim	AIML[7] e Técnicas de RI
[Prestes 2011]	genérico	sim	Análise de Padrões
[Arrigo et al. 2014]	genérico	não	Análise de Padrões

resumos de artigos científicos do MEDLINE, PubMed, eMedicine e Wikipédia. O sistema integra UMLS [Bodenreider 2004, Brin 1999] na expansão dos termos da consulta do usuário, técnicas de PLN [Liddy 2001], RI [Baeza-Yates and Ribeiro-Neto 1999] e EI [Cowie and Lehnert 1996]. Segundo os autores, o sistema apresentou uma boa habilidade na solução de questões longas e complexas [Cao et al. 2011]. O segundo sistema do domínio médico é o MEANS, apoiado por abordagens semânticas de PLN e Web Semântica. As abordagens semânticas são utilizadas para análise dos dados em dois níveis de complexidade: busca por entidades médicas (drogas, sintomas e doenças) e análise em nível de relacionamento (tratamentos, prevenções e causas). As questões são focadas nos tipos factuais e booleanas [Ben Abacha and Zweigenbaum 2015].

No Brasil, o desenvolvimento de sistemas de PR é pouco explorado. Um dos sistemas encontrados foi desenvolvido para domínios de informações genéricas. Segundo os autores, o sistema possui uma arquitetura flexível que independe do conteúdo da pergunta inserida [Amorim et al. 2012]. Para que o sistema suporte esta abrangência, os autores utilizaram um banco de ontologias junto a uma gama de softwares de apoio. Caso a resposta não seja encontrada neste banco de dados, o sistema busca a resposta na Web.

Neste trabalho apresentado, o autor comparou os seguintes sistemas de PR, com foco no módulo de processamento da resposta levando em consideração a complexidade da linguagem a qual foram criados: (i) um sistema para o idioma português, (ii) um para o idioma inglês e (iii) um em idioma português, porém estendido para o inglês. O autor verificou o uso de regras semânticas e sintáticas e técnicas PLN [Prestes 2011]. Por fim, o último sistema de PR apresentado utiliza a Web para criar um corpus anotado de informações em Português. O sistema faz análise semântica da pergunta e utiliza métodos de aprendizado automático de padrões para classificação da informação e para a análise da precisão das respostas [Arrigo et al. 2014]. A Tabela 2 apresenta uma organização sistematizada das arquiteturas encontradas na literatura. As principais semelhanças entre o QASF e as demais arquiteturas são o uso de artefatos linguísticos e SVM, quando o método de classificação é por AM.

O QASF quando comparado com os demais sistemas de informação do tipo PR possui alguns diferenciais. O QASF foi inicialmente projetado para atuar no domínio médico, cuja busca por informações clínicas relevantes pode ser mais complexa quando

comparada a outros domínios da informação. Por essa razão, o módulo de recuperação de documentos foi construído baseado em um sistema especialista em recuperação de informações médicas, o SisViDAS. A arquitetura do sistema QASF é ampla e pode ser estendida para vários outros domínios de informação, variando apenas os artefatos linguísticos como as ontologias. Outro diferencial é que o sistema está sendo desenvolvido para processar perguntas com diferentes níveis de complexidade, ou seja, perguntas que exigem diferentes níveis de processamento do QASF para encontrar a resposta mais adequada.

5. Conclusões

No ambiente médico, buscar informações relevantes em artigos científicos é uma tarefa que exige tempo e experiência dos profissionais. Portanto, o objetivo deste trabalho é apresentar uma arquitetura de um sistema do tipo Pergunta e Respostas (PR) que auxilie profissionais da área da saúde na busca rápida por respostas. Atualmente o sistema está focado em doenças crônicas e os módulos de processamento da questão, documento e resposta estão em fase final de avaliações. No submódulo de "Análise da Questão", avaliações apontam que SVM apresenta melhores resultados na classificação de perguntas. Como trabalhos futuros, pretende-se: (i) avaliar o QASF em outros domínios de informação como, por exemplo, o câncer de tireoide e (ii) transformar o QASF em aplicação composta por *interfaces* de usuário web e para dispositivos móveis que facilite a interação com o sistema.

6. Agradecimento

Os autores gostariam de agradecer a CAPES pelo apoio financeiro concedido para o desenvolvimento do projeto.

Referências

Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *Int. Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3):211–220.

Amorim, M. T. C. F. d., Cury, D., and Menezes, C. S. (2012). Um Sistema Inteligênte Baseado em Ontologia para Apoio ao esclarecimento de Dúvidas.

Arrigo, A. J. S., Silva, E. G., Martins, H. P., and Silva, P. P. (2014). Desenvolvimento de um Sistema de Pergunta e Resposta Baseado em Corpus. In *14o Congresso Nacional de Iniciação Científica (CONIC-SEMESP)*, pages 1–6, São Paulo, SP.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM Press New York, 1nd edition.

Barker, D. J. P. (2001). Fetal and infant origins of adult disease. *Monatsschrift Kinderheilkunde*, 149(1):S2–S6.

Ben Abacha, A. and Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information Processing & Management*, 51(5):570–594.

Bhat, S., Gijo, E., and Jnanesh, N. (2016). Productivity and performance improvement in the medical records department of a hospital: An application of lean six sigma. *International Journal of Productivity and Performance Management*, 65(1):98–125.

Bilotti, M. W. and Nyberg, E. (2006). Evaluation for scenario question answering systems. In *Proc. of the Int. Conference on Language Resources and Evaluation*, pages 1–6.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Brin, S. (1999). Extracting patterns and relations from the world wide web. In *Selected Papers from the International Workshop on The World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK, UK. Springer-Verlag.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., Ely, J., and Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–88.

Cowie, J. and Lehnert, W. (1996). Information extraction. *Comm. of the ACM*, 39(1):80–91.

Er, N. P. and Cicekli, I. (2013). A Factoid Question Answering System Using Answer Pattern Matching. In *Int. Joint Conf. on Natural Language Processing*, pages 854–858.

Gupta, P. and Gupta, V. (2012). A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4):1–8.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412 – 5434.

Liddy, E. D. (2001). Natural Language Processing. In Decker, M., editor, *In Encyclopedia of Library and Information Science*, pages 1–15. New York, New York, USA.

Lin, J. and Demner-Fushman, D. (2005). Automatically evaluating answers to definition questions. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 931–938.

Miyake, K., Hirasawa, T., Koide, T., and Kubota, T. (2012). *Neurodegenerative Diseases*, chapter Epigenetics in Autism and Other Neurodevelopmental Diseases, pages 91–98.

Monz, C. (2003). *From document retrieval to question answering*. Institute for Logic, Language and Computation.

Pollettini, J. T., Baranauskas, J. A., Ruiz, E. S., da Graça Pimentel, M., and Macedo, A. A. (2014). Surveillance for the prevention of chronic diseases through information association. *BMC Medical Genomics*, 7(1):1–11.

Prestes, K. V. (2011). Avaliação de métodos de seleção da resposta de um sistema de perguntas e respostas. Technical report.

Rzhetsky, A., Seringhaus, M., and Gerstein, M. (2009). Getting Started in Text Mining: Part Two. *PLoS Comput Biol*, 5(7):e1000411+.

Shortliffe, E. H. and Cimino, J. J. (2013). *Biomedical informatics: computer applications in health care and biomedicine*. Springer Science & Business Media.

Yen, S.-J., Wu, Y.-C., Yang, J.-C., Lee, Y.-S., Lee, C.-J., and Liu, J.-J. (2013). A support vector machine-based context-ranking model for question answering. *Information Sciences*, 224:77 – 87.