Extração de Entidades e Relações para um Sistema de Question Answering no Domínio da Saúde

Nicolas Eymael da Silva, Dante Augusto Couto Barone, Eduardo Gabriel Cortes[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

Abstract. *In the health area, scientific articles are a rich source of information for professionals in the field. However, with the growth in the volume of publications, reading articles has become an arduous task that requires time and willingness from health professionals. In order to help users reduce the time required in the search process, QA systems can provide a quick answer to a question asked in natural language. The goal of this work is the development of computational models that are able to extract entities and relationships from text documents in Portuguese related to the health area. These models will be tested, evaluated and, later, selected to build a graph database that will be used to feed a diagrammatic QA system.*

Resumo. *Na área de saúde, os artigos científicos são uma fonte de informação rica para os profissionais da área. No entanto, com o crescimento no volume de publicações, a leitura de artigos se tornou uma tarefa árdua que exige tempo e disposição dos profissionais da saúde. Com o intuito de auxiliar os usuários a reduzir o tempo necessário no processo de busca, os sistemas de QA podem fornecer uma resposta rápida a uma pergunta feita em linguagem natural. O objetivo desse trabalho é o desenvolvimento de modelos computacionais que sejam capazes de extrair entidades e relacionamentos a partir de documentos de texto em português relacionados a área da saúde. Esse modelos serão testados, avaliados e, posteriormente, selecionados para construir uma base de dados baseada em grafos que irá ser usada para alimentar um sistema de QA diagramático.*

1. Introdução

A Biblioteca Virtual em Saúde (BVS) é um serviço do Ministério da Saúde no qual são publicadas as informações bibliográficas relacionadas a área de ciências da saúde. Tal serviço já consta com mais de 25 milhões de artigos em seu catálogo, majoritariamente em língua inglesa. Além disso, a publicação científica global está crescendo a uma taxa em torno de 3% ao ano. Isso indica que o volume de publicações aproximadamente dobra a cada 24 anos [Bornmann and Mutz 2015].

Na área de saúde, os artigos científicos são uma fonte de informação rica para os profissionais da área. No entanto, diante desse vasto acervo bibliográfico, ler artigos científicos se tornou uma tarefa árdua que exige tempo e disposição e é natural que os profissionais da saúde apresentem dificuldades ao obter as informações desejadas [Almansa 2016].

Nesse contexto, de forma a melhorar o aproveitamento das informações contidas nos artigos científicos, percebe-se a necessidade de adotar ferramentas que auxiliam no

processo de busca de informações específicas. Por exemplo, um sistema de *Question Answering* (QA) voltado especificamente para o domínio da saúde. Os sistemas de QA são ferramentas que podem auxiliar usuários a reduzir o tempo necessário no processo de busca, já que são capazes de fornecer uma resposta rápida e precisa a uma pergunta feita em linguagem natural [Abacha and Zweigenbaum 2015].

Antes que um sistema de QA consiga responder perguntas corretamente, ele precisa ser treinado com uma base de conhecimento. Essa base pode ser estruturada (baseada em grafos) ou não estruturada (baseada em textos) [Moschitti et al. 2017]. Enquanto que os documentos de texto são mais informativos, os grafos de conhecimento são mais eficientes pois eles já fornecem os dados mais relevantes que foram extraídos desses textos.

A construção de um grafo de conhecimento se dá através do processamento de documentos para extrair as entidades e relacionamentos presentes nos textos. Cada entidade irá ser representada como um nó no grafo e os relacionamentos serão as arestas que ligam pares de nós. Por exemplo, na frase ”Obama nasceu no Havaí”, o relacionamento ”nascer em”é o que liga a entidade ”Obama”a entidade ”Havaí”.

Portanto, o objetivo dessa pesquisa é o desenvolvimento de modelos computacionais que sejam capazes de extrair entidades e relações a partir de dados não estruturados da área da saúde (comumente textos), preferencialmente em português. Esse modelos serão testados para a alimentação de uma base de dados em grafos que será utilizada por um sistema de QA.

O restante desse artigo está organizado da seguinte maneira. Na seção 2, é apresentado o referencial teórico a respeito da construção de grafos de conhecimento, assim como sua relação com os sistemas de QA. Na seção 3, é apresentada a metodologia adotada nessa pesquisa e quais os próximos passos do projeto. Na seção 4, é apresentado o cronograma do projeto. Por fim, a seção 5 conclui o artigo.

2. Referencial Teórico

De modo a fornecer o embasamento teórico necessário, essa seção apresenta as áreas da computação diretamente relacionadas aos grafos de conhecimento e sistemas de QA [Côrtes 2019].

Primeiramente, processamento de linguagem natural (PLN) é uma área fundamental da inteligência artificial, cujo objetivo é que os computadores sejam capazes de compreender textos como um humano entenderia [Indurkhya and Damerau 2010]. Por meio de PLN, é possível construir bases estruturadas ao realizar a ligação das entidades nomeadas [Han and Sun 2011] e a extração das relações semânticas [Sarawagi 2008] de um texto.

Um tipo de base estruturada que vem ganhando destaque nos últimos tempos é o grafo de conhecimento (do inglês *Knowledge Graph*, KG) [Wang et al. 2017]. Um KG representa uma coleção de descrições de entidades interligadas em rede, em que essas descrições possuem uma alta expressividade. Essa expressividade se dá pois as descrições possuem semânticas formais baseadas em ontologias, o que garante um melhor gerenciamento dos dados [de Lima and de Carvalho 2005].

À medida que a quantidade de dados disponíveis continua crescendo, houve um aumento na busca por meios de melhor aproveitar esse crescente volume de informação

através de interfaces intuitivas. Dessa forma, os sistemas de QA baseados em KGs tem recebido atenção de pesquisadores por serem capazes de retornar respostas diretas e precisas a perguntas possivelmente complexas [Unger et al. 2014].

2.1. Ligação de Entidades (EL)

A ligação de entidades (do inglês *Entity Linking*, EL) se refere a tarefa de reconhecer e desambiguar entidades mencionadas em dados não estruturados para suas entidades correspondentes em um KG [Shen et al. 2014]. A classe de EL abordada nesse artigo é a fim-a-fim, ou seja, o processamento é iniciado a partir de um texto puro.

A primeira etapa da EL fim-a-fim é o reconhecimento das entidades nomeadas (do inglês *Named Entity Recognition*, NER). Essa etapa é responsável por identificar as ocorrências de entidades nomeadas em um texto e classificá-las de acordo com categorias pré-definidas [Nadeau and Sekine 2007].

A Figura 1 mostra o resultado do NER aplicado a um parágrafo de um texto. O NER é capaz de identificar, por exemplo, que *"Sebastian Thrun"* é uma entidade do tipo Pessoa e que *"Google"* é uma entidade do tipo Organização. No entanto, o NER não é capaz de concluir que as entidades *"Sebastian Thrun"* e *"Thrun"* na verdade se referem à mesma pessoa.

Figura 1. NER aplicado a um parágrafo de um texto.

De forma a realizar essa conexão entre a entidade identificada no texto e a entidade presente no KG, é necessário executar a segunda etapa da EL, chamada de desambiguação de entidades nomeadas (do inglês *Named Entity Disambiguation*, NED) [Hoffart et al. 2011]. O NED é responsável por vincular a entidade corretamente ao KG, tanto nos casos em que uma palavra possui vários significados distintos quanto nos casos em que palavras distintas possuem o mesmo significado (o que ocorre com *"Thrun"* no exemplo).

A Figura 2 mostra o resultado do NED aplicado a uma frase. A entidade *"Paris"* possui mais de uma correspondência no KG, pois pode estar relacionada tanto à capital da França quanto a uma pequena cidade do Arkansas, EUA. Do mesmo modo, a entidade *"France"* também pode se referir a seleção de futebol francês, por exemplo.

É de responsabilidade do NED realizar essa diferenciação dos significados e relacionar as entidades corretamente às entidades correspondentes no KG. Nesse caso, "Paris (a cidade da França) é a capital da França (país da Europa)".

A maior parte dos trabalhos de EL presentes na literatura abordam o NER e o NED de forma independente. Dessa forma, é possível obter resultados surpreendentes com taxas de até 99% de acurácia [Raiman and Raiman 2018].

Figura 2. NED aplicado a uma frase.

No entanto, o desafio final é um modelo de EL fim-a-fim capaz de unir as duas etapas e suas diferentes naturezas. Alguns trabalhos já abordaram a modelagem fim-a-fim [Kolitsas et al. 2018, van Hulst et al. 2020, Piccinno and Ferragina 2014] e obtiveram bons resultados, mas ainda não se comparam às abordagens com métodos independentes.

2.2. Extração de Relacionamentos (RE)

Após identificar e vincular as entidades nomeadas do texto, é necessário descobrir como tais entidades relacionam-se entre si. A extração de relacionamentos (do inglês *Relation Extraction*, RE) é a tarefa de predizer os atributos e relações entre as entidades em uma sentença [Huang and Wang 2017].

Os relacionamentos geralmente ocorrem entre duas ou mais entidades não necessariamente do mesmo tipo. Uma entidade do tipo Pessoa, por exemplo, pode ter uma relação semântica com um Local (como a relação ”nasceu em”) ou até mesmo com outra Pessoa (como a relação ”é filha de”).

Apesar da RE ser um componente crucial na construção de KGs, ela é considerada uma das tarefas mais difíceis na área de PLN [Lin et al. 2016]. Além de detectar se existe ou não algum tipo de relação (geralmente binária) entre entidades, ainda é preciso determinar em que classe aquela relação se encaixa.

A Figura 3 mostra o resultado da ER em uma sentença. A entidade *”Bill Gates”* possui um relacionamento de “nascido na data”com a entidade “28 de Outubro de 1955”. Além disso, a mesma entidade *“Bill Gates”* também possui outro relacionamento, dessa vez de “nascido no local”, com a entidade *”Seattle”*. Apesar de ”28 de Outubro de 1955”e *“Seattle”* estarem indiretamente ligadas através do *“Bill Gates”*, as duas entidades não possuem nenhum relacionamento entre si.

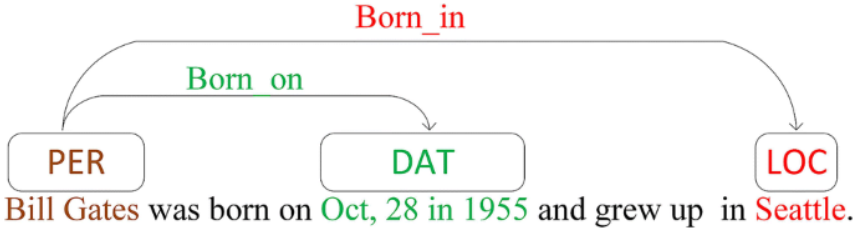

Figura 3. ER aplicado a uma sentença.

A tarefa de RE pode ser abordada de diferentes maneiras. Existem métodos supervisionados que utilizam anotações prévias, métodos não supervisionados que são ba-

seados em padrões genéricos de extração, ou até mesmo métodos semi-supervisionados que aplica conceitos dos outros dois métodos. Infelizmente, um problema comum a todos os métodos é a escassez de modelos voltados à Língua Portuguesa [de Abreu et al. 2013].

Em contrapartida, já existe uma variedade de trabalhos em inglês utilizando métodos e conjuntos de dados distintos. Modelos utilizando extração distantemente supervisionada aplicados em textos do *New York Times* são capazes de atingir por volta de 80 a 85% de precisão [Xu and Barbosa 2019, Wu et al. 2019, Ye and Ling 2019].

2.3. Grafo de Conhecimento (KG)

Após o processamento das técnicas de EL e RE em conjunto sobre textos, é possível construir um KG para agregar os dados coletados. Basicamente, um KG é uma grande rede de entidades composta por tipos, atributos e relacionamentos semânticos entre essas entidades [Kroetsch and Weikum 2016].

A Figura 4 mostra um KG com entidades refente a filmes de ficção científica. Outra característica do KG é ser expansível, ou seja, é possível adicionar uma nova entidade (o ator *"Ewan McGregor"*, por exemplo) e relacioná-la com as outras entidades já existentes no grafo (como um relacionamento de "interpretou"com a entidade *"Obi-Wan Kenobi"*). Além disso, esse processo de expansão pode até mesmo ser automatizado [Yoo and Jeong 2020].

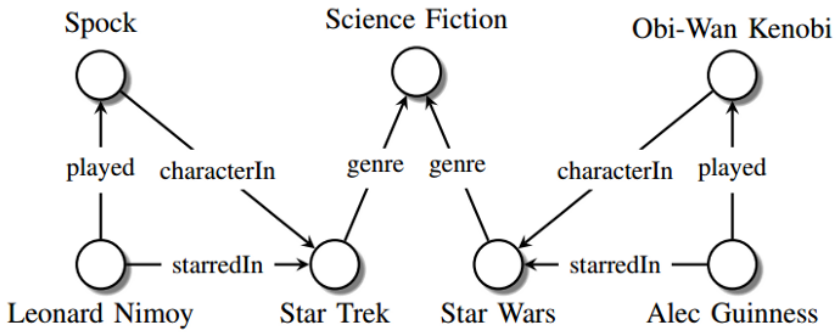

Figura 4. Inferência de relação entre entidades.

A unidade elementar de um KG é a tripla sujeito-predicado-objeto [Nickel et al. 2015]. Em um grafo qualquer, cada tripla define uma conexão entre dois nós. No caso de um KG, as conexões se tratam dos relacionamentos e os nós são as entidades.

Além disso, o KG também é capaz de inferir relações entre entidades [Liu et al. 2016]. Na Figura 5 é apresentado um exemplo de inferência de relação entre as entidades "eu"e "avô". Mesmo que no grafo original as duas entidades não estejam diretamente ligadas, é possível realizar essa conexão através da entidade "pai".

2.4. Sistema de QA baseado em KG

Uma das aplicações mais usuais para um KG é um sistema de QA [Huang et al. 2019]. O KG possui uma abundância de informações e o sistema de QA proporciona ao usuário um acesso mais fácil e eficiente às informações desejadas, de modo que ele não necessite conhecer a estrutura dos dados para navegar pelo grafo.

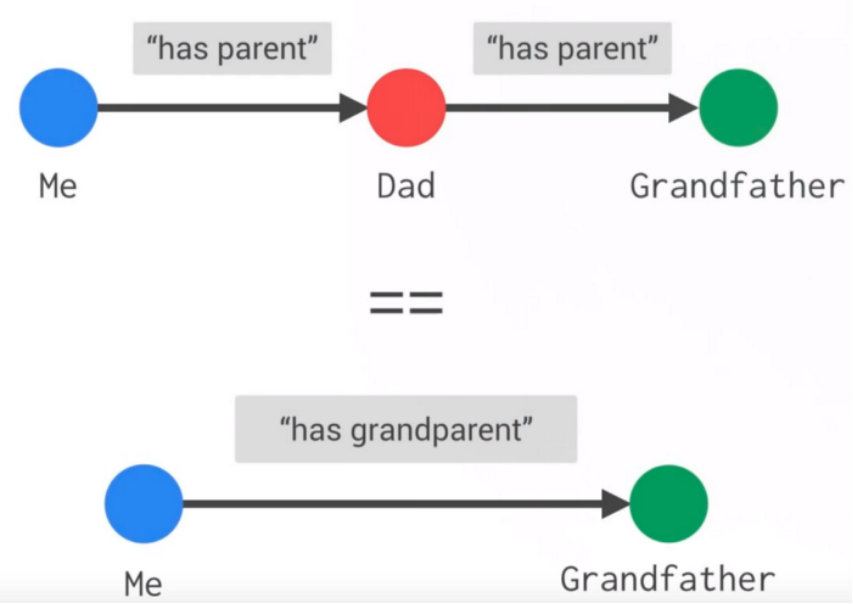

Figura 5. KG com entidades relacionadas a filmes.

A tarefa do sistema de QA baseado em KG (denominado como KGQA) envolve responder uma pergunta feita em linguagem natural utilizando as informações armazenadas no KG. O primeiro passo é traduzir a pergunta de entrada para uma linguagem de consulta formal (como SPARQL) e só depois executar essa consulta no KG para obter a resposta [Chakraborty et al. 2019].

Essa tradução geralmente é feita através de uma análise semântica. No entanto, se a análise da pergunta possui erros, consequentemente será gerada uma consulta falha e a resposta retornada pelo KGQA não necessariamente estará correta. Por isso, muitos trabalhos dão ênfase no processo de tradução da pergunta e propõem métodos distintos, utilizando modelos de questões ou até mesmo outros grafos [Zheng et al. 2018, Zheng and Zhang 2019, Yih et al. 2015].

Uma variação interessante do KGQA é a sua abordagem visual, também chamado de sistema de QA diagramático (do inglês *Diagrammatic Question Answering*, DQA) [Mouromtsev et al. 2018]. Esse sistema auxilia o usuário a navegar pelo KG e a detectar as propriedades e relações mais relevantes de cada entidade.

A Figura 6 mostra o resultado do DQA para a entidade UFRGS e sua localização (cidade, estado e país). Além dos relacionamentos de ”localizado em”, o sistema também apresenta outros relacionamentos como ”capital de”e ”país”.

Por não ser um sistema automatizado, o DQA demanda mais interação manual com o usuário. Apesar desse empecilho, os resultados indicam que o sistema apresenta um desempenho melhor do que o KGQA tradicional [Wohlgenannt et al. 2019]. Além disso, as duas abordagens podem ser utilizadas em combinação de maneira complementar.

3. Metodologia

Como já foi dito, o propósito desse trabalho é a implementação de modelos capazes de extrair entidades e relacionamentos de textos em português relacionados ao domínio

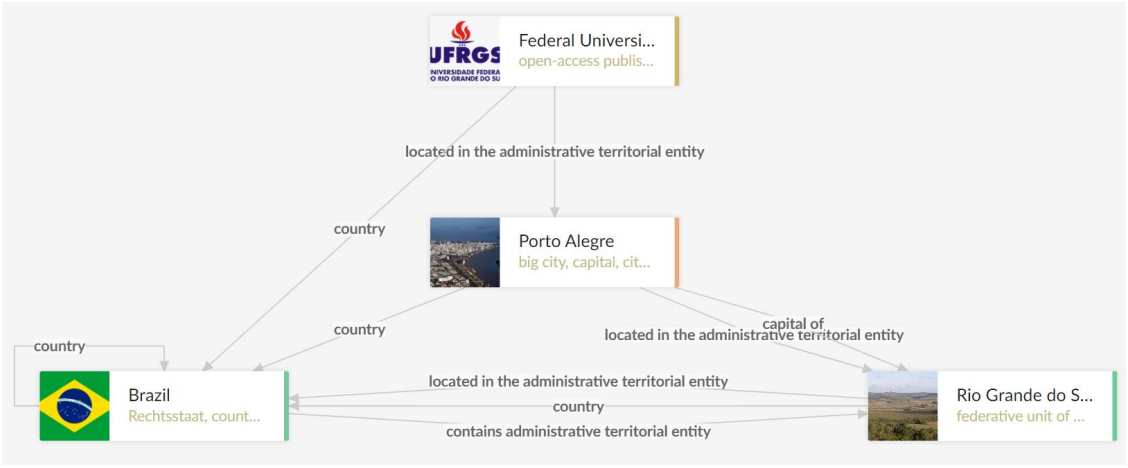

Figura 6. Resultado do DQA para UFRGS e sua localização.

da saúde. Com essa finalidade, o trabalho foi dividido em etapas para uma melhor organização.

O primeiro passo é a pesquisa sobre o tema e o estudo de trabalhos relacionados. O próximo passo é a seleção dos conjuntos de dados a serem utilizados no treinamento e teste dos modelos. Em seguida, será realizada a implementação dos modelos escolhidos na etapa de pesquisa. Esses modelos estão relacionados aos processos de construção de bases estruturadas a partir de textos.

Depois será realizado o treinamento, teste e avaliação dos modelos utilizando os conjuntos de dados com o intuito de selecionar aqueles que são mais eficientes. Os melhores modelos serão selecionados para extrair entidades e relacionamentos de textos da área da saúde. Com isso, será criada uma base de dados que irá alimentar um sistema de DQA. Por fim, o sistema será avaliado junto a especialistas na saúde.

A primeira subseção irá descrever as características dos conjuntos de dados escolhidos. A seguinte irá abordar o treinamento e teste de modelos de extração de entidades e relacionamentos. Por fim, a última irá tratar sobre como os modelos selecionados serão utilizados em um sistema de QA.

3.1. Conjuntos de Dados

Dois conjuntos de dados serão usados para o treinamento dos modelos. Esses conjuntos foram escolhidos por serem totalmente em português, o idioma alvo do nosso modelo, e por já possuírem dados anotados, o que facilita a etapa de testes.

O primeiro conjunto de dados escolhido consta com 98.023 exemplos em português extraídos da Wikipédia e DBPédia. A construção desse conjunto de dados se deu da seguinte forma [Batista et al. 2013]:

1. Recolhem-se da DBPédia todas as relações expressas entre entidades correspondentes a pessoas, locais ou organizações;
2. Para cada relação entre um par de entidades, analisa-se o texto dos dois artigos correspondentes da Wikipédia em português;
3. O texto dos artigos é segmentado em sentenças;

4. As sentenças são filtradas, de modo a manter apenas aquelas em que as duas entidades estão presentes;
5. As sentenças que resultam da etapa de filtragem são mantidas como exemplos de relação semântica.

Entre os exemplos, encontram-se 3 tipos de entidades: locais (141.028 instâncias), pessoas (38.903 instâncias) e organizações (16.115 instâncias). Além disso, os relacionamentos são divididos em 10 tipos distintos que podem ser visualizados na Tabela 1.

Relacionamento	Ocorrências
locatedInArea	46.864
keyPerson	392
origin	26.236
deathOrBurialPlace	7.146
successor	567
partner	190
parent	298
influencedBy	154
partOf	5.520
other	10.656

Tabela 1. Tipos de relacionamentos presentes no conjunto de dados da DBPédia.

O outro conjunto de dados escolhido é o Segundo HAREM, organizado pela equipe da Linguateca [Mota and Santos 2008]. A coleção dourada do Segundo HAREM é composta por 129 documentos em português, cujas entidades e relacionamentos foram manualmente anotadas e classificadas.

As entidades da coleção são divididas em 10 categorias diferentes e cada categoria pode ter tipos e subtipos [Carvalho and Freitas 2008]. O conjunto possui aproximadamente 7.800 entidades e 56.900 relacionamentos entre elas. O número de instâncias de cada entidade e o número de ocorrências de alguns relacionamentos podem ser visualizados nas Tabelas 2 e 3, respectivamente (valores não exatos). Alguns relacionamentos possuem relacionamento inverso.

3.2. Experimentos

De modo a construir uma base estruturada referente ao domínio da saúde, serão realizados experimentos voltados tanto para o processo de EL quanto para o processo de RE. Para cada etapa, serão utilizados diferentes modelos que foram obtidos durante a atividade de pesquisa do projeto.

Esses modelos serão testados com os conjuntos de dados escolhidos e comparados entre si, com o intuito de descobrir qual modelo é o mais eficiente. A técnica de validação cruzada (do inglês *cross-validation*, CV) será utilizada para avaliar cada modelo [Pedregosa et al. 2011].

O CV tem como principal objetivo evitar problemas de aleatoriedade dos dados e, com isso, obter um resultado mais robusto. A técnica consiste em repartir os dados em grupos e testar o modelo em várias iterações. A cada iteração, um grupo de dados

Entidade	Instâncias
PESSOA	2.091
LOCAL	1.453
TEMPO	1.199
ORGANIZAÇÃO	1.080
OBRA	497
ABSTRAÇÃO	382
VALOR	352
COISA	345
ACONTECIMENTO	337
OUTRO	81

Tabela 2. Tipos de entidades presentes no conjunto de dados do HAREM.

Relacionamento	Ocorrências
ident	17.560
vinculo_inst	5.766
inclui (incluido)	5.020
ocorre_em (sede_de)	4.138
obra_de (autor_de)	1.755
relacao_familiar	1.296
natural_de (local_nascimento_de)	1.048
propriedade_de (proprietario_de)	953
produtor_de (produzido_por)	812
participante_em (ter_participacao_de)	636

Tabela 3. Exemplos de relacionamentos do conjunto de dados do HAREM.

será escolhido como o grupo de teste enquanto que o restante dos grupos será usado no treinamento do modelo. Dessa forma, evita-se o sobre-ajuste (do inglês *overfitting*) do modelo e se assegura a sua capacidade de generalização para um novo conjunto de dados.

As métricas que serão utilizadas na comparação dos modelos são a precisão, a revocação e o F1. As fórmulas de cada uma dessas métricas podem ser visualizadas na Figura 7. A Tabela 4 mostra os 4 possíveis resultados de um sistema de predição através de uma matriz de confusão.

A precisão estima quantos resultados são verdadeiro positivo dentre todas as predições positivas (boa métrica para modelos em que cada falso positivo é importante). A revocação estima quantos resultados são verdadeiro positivo dentre todos os valores realmente positivos (boa métrica para modelos em que cada falso negativo é importante). Finalmente, o F1 é uma boa métrica quando é necessário um equilíbrio entre precisão e revocação em um conjunto de dados com uma distribuição desigual.

3.3. Avaliação do modelo com especialista da saúde

Após os experimentos, os modelos mais eficientes de cada processo serão escolhidos para a próxima etapa do projeto. Esse etapa consiste em aplicar tais modelos sobre documentos

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figura 7. Fórmulas para precisão, revocação e F1, respectivamente.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Tabela 4. Matriz de confusão.

de texto em português relacionados ao domínio da saúde. Dessa forma, serão extraídas as entidades e relacionamentos que servirão para construir uma base de dados.

Essa base de dados será posteriormente utilizada para alimentar a ferramenta de DQA ”metaphactory”[Mouromtsev et al. 2018]. Essa ferramenta, desenvolvida na Universidade de São Petersburgo (*ITMO University*), permite a visualização do grafo de conhecimento através de uma interface gráfica com o usuário.

Como já foi dito, um DQA requer uma maior interação do usuário do que um sistema de QA convencional. Portanto, a ferramenta será manuseada por especialistas da área da saúde para que seja realizada uma melhor avaliação do modelo.

Por último, iremos propor um aprimoramento da ferramenta de DQA. Como foi dito anteriormente, é possível combinar o DQA com um sistema de QA tradicional. Nesse caso, o usuário poderia acessar o sistema através de uma pergunta em linguagem natural e inspecionar o resultado da consulta através da interface visual do grafo.

4. Cronograma

O trabalho proposto está previsto para ser concluído em um período de até 6 meses. Portanto, as atividades planejadas para o próximo semestre são as seguintes:

1. Estudo dos assuntos e trabalhos relacionados.
2. Implementação dos experimentos com diferentes modelos para cada tarefa.
3. Análise de resultados e melhorias.
4. Alimentação da base de conhecimento baseada em grafo.
5. Inserção no sistema de visualização de base de dados em grafo.
6. Avaliação com especialista na saúde.
7. Escrita do Trabalho de Conclusão.
8. Escrita e submissão de Artigo.

A Tabela 5 apresenta o cronograma do projeto com os meses previstos para a realização de cada uma dessas atividades.

Atividade	M1	M2	M3	M4	M5	M6
1	X					
2		X				
3		X	X			
4			X	X		
5			X	X		
6				X	X	
7	X				X	X
8						X

Tabela 5. Cronograma de Atividades.

5. Conclusão

Assim, conforme o que foi pesquisado, constatou-se que os sistemas de QA são ferramentas muito adequadas para reduzir o tempo de busca de informações. Como essas ferramentas geralmente são voltadas para a língua inglesa, propôs-se o desenvolvimento de modelos direcionados especificamente para documentos em português do domínio da saúde.

Até o momento, foi realizada uma pesquisa sobre os principais conceitos da extração de entidades e relacionamentos e o estudo de trabalhos relacionados. Além disso, já foi feita a seleção dos conjuntos de dados a serem utilizados e o planejamento do próximo semestre.

Para o trabalho futuro, a meta é implementar e treinar diferentes modelos de extração e escolher os mais eficientes (de acordo com métricas já determinadas) para construir a base estruturada que alimentará o DQA. Outra proposta ainda é o desenvolvimento de um sistema que une o DQA com o QA convencional para fornecer uma ferramenta mais poderosa para os usuários.

Referências

Abacha, A. B. and Zweigenbaum, P. (2015). Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Almansa, L. F. (2016). *Uma arquitetura de question-answering instanciada no domínio de doenças crônicas*. PhD thesis, Universidade de São Paulo.

Batista, D. S., Forte, D., Silva, R., Martins, B., and Silva, M. (2013). Extracçao de relaçoes semânticas de textos em português explorando a dbpédia e a wikipédia. *linguamatica*, 5(1):41–57.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222.

Carvalho, P. and Freitas, C. (2008). Apêndice e: Exemplário do segundo harem. *quot; In Cristina Mota; Diana Santos (ed) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM Linguateca 2008*.

Chakraborty, N., Lukovnikov, D., Maheshwari, G., Trivedi, P., Lehmann, J., and Fischer, A. (2019). Introduction to neural network based approaches for question answering over knowledge graphs. *arXiv preprint arXiv:1907.09361*.

Côrtes, E. G. (2019). Quando, onde, quem, o que ou por que? um modelo híbrido de classificação de perguntas para sistemas de question answering.

de Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013). A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571.

de Lima, J. C. and de Carvalho, C. L. (2005). Ontologias-owl (web ontology language). Technical report, Technical report, Universidade Federal de Goiás.

Han, X. and Sun, L. (2011). A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 945–954.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792.

Huang, X., Zhang, J., Li, D., and Li, P. (2019). Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 105–113.

Huang, Y. Y. and Wang, W. Y. (2017). Deep residual learning for weakly-supervised relation extraction. *arXiv preprint arXiv:1707.08866*.

Indurkhya, N. and Damerau, F. J. (2010). *Handbook of natural language processing*, volume 2. CRC Press.

Kolitsas, N., Ganea, O.-E., and Hofmann, T. (2018). End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.

Kroetsch, M. and Weikum, G. (2016). Special issue on knowledge graphs. *Journal of Web Semantics*, 37(38):53–54.

Lin, Y., Shen, S., Liu, Z., Luan, H., and Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Liu, Q., Jiang, L., Han, M., Liu, Y., and Qin, Z. (2016). Hierarchical random walk inference in knowledge graphs. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 445–454.

Moschitti, A., Tymoshenko, K., Alexopoulos, P., Walker, A., Nicosia, M., Vetere, G., Faraotti, A., Monti, M., Pan, J. Z., Wu, H., et al. (2017). Question answering and knowledge graphs. In *Exploiting Linked Data and Knowledge Graphs in Large Organisations*, pages 181–212. Springer.

Mota, C. and Santos, D. (2008). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem.

Mouromtsev, D., Wohlgenannt, G., Haase, P., Pavlov, D., Emelyanov, Y., and Morozov, A. (2018). A diagrammatic approach for visual question answering over knowledge graphs. In *European Semantic Web Conference*, pages 34–39. Springer.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Piccinno, F. and Ferragina, P. (2014). From tagme to wat: a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, pages 55–62.

Raiman, J. and Raiman, O. (2018). Deeptype: multilingual entity linking by neural type system evolution. *arXiv preprint arXiv:1802.01021*.

Sarawagi, S. (2008). *Information extraction*. Now Publishers Inc.

Shen, W., Wang, J., and Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Unger, C., Freitas, A., and Cimiano, P. (2014). An introduction to question answering over linked data. In *Reasoning Web International Summer School*, pages 100–140. Springer.

van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., and de Vries, A. P. (2020). Rel: An entity linker standing on the shoulders of giants. *arXiv preprint arXiv:2006.01969*.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Wohlgenannt, G., Mouromtsev, D., Pavlov, D., Emelyanov, Y., and Morozov, A. (2019). A comparative evaluation of visual and natural language question answering over linked data. *arXiv preprint arXiv:1907.08501*.

Wu, S., Fan, K., and Zhang, Q. (2019). Improving distantly supervised relation extraction with neural noise converter and conditional optimal selector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7273–7280.

Xu, P. and Barbosa, D. (2019). Connecting language and knowledge with heterogeneous representations for neural relation extraction. *arXiv preprint arXiv:1903.10126*.

Ye, Z.-X. and Ling, Z.-H. (2019). Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv preprint arXiv:1904.00143*.

Yih, S. W.-t., Chang, M.-W., He, X., and Gao, J. (2015). Semantic parsing via staged query graph generation: Question answering with knowledge base.

Yoo, S. and Jeong, O. (2020). Automating the expansion of a knowledge graph. *Expert Systems with Applications*, 141:112965.

Zheng, W., Yu, J. X., Zou, L., and Cheng, H. (2018). Question answering over knowledge graphs: question understanding via template decomposition. *Proceedings of the VLDB Endowment*, 11(11):1373–1386.

Zheng, W. and Zhang, M. (2019). Question answering over knowledge graphs via structural query patterns. *arXiv preprint arXiv:1910.09760*.