Athena!: um protótipo de um sistema de perguntas e respostas a partir de base de dados abertas e conectadas

Augusto Lopes da Silva[1], **Sandro José Rigo**[1],
Angelita Barbosa Nunes[2], **Jorge Barbosa**[1]

[1]Programa de Pós-Graduação em Computação Aplicada (PIPCA)
Universidade do Vale do Rio dos Sinos (UNISINOS) – São Leopoldo, RS – Brasil

[2]Departamento Acadêmico de Informática (DAINF)
Universidade Tecnológica Federal do Paraná – Ponta Grossa, PR – Brasil

{augustols,angelita.nunes}@outlook.com, {rigo,jbarbosa}@unisinos.br

Abstract. *This article presents a system prototype for a question answering system exploring knowledge from linked open data. The prototype was developed to support teachers and students in their regular activities. Students answer questions about topics from a knowledge domain and, for each correct answer, they score points in a ranking system. To validate the prototype, middle and high-school classes students and their teachers were invited to test and answer a simple survey about it. Based on the positive answers, this study suggests that linked data can be useful for educational purposes when combined with question answering systems.*

Resumo. *Este artigo apresenta um protótipo para um sistema de perguntas e respostas, que tem como objetivo explorar o conhecimento disponível em bases de dados abertas e conectados. Ele foi desenvolvido para auxiliar professores e alunos em suas atividades de estudo. Estudantes respondem questões sobre tópicos de um domínio de conhecimento, e, para cada resposta correta, ganham pontos dentro do ranking do sistema. Para validação do protótipo, estudantes e seus professores foram convidados a testar o sistema e responder algumas perguntas nele. Baseado nas respostas positivas, este estudo sugere que dados abertos e ligados podem ser úteis para propósitos educacionais quando combinados com sistemas de perguntas e respostas.*

1. Introdução

Houve um aumento exponencial na quantidade de dados gerados, publicados e compartilhados na Internet [Mesinovic 2013]. No entanto, esse conteúdo, na maioria das vezes, não é estruturado a ponto de ser compreendido por máquinas.

Grupos de pesquisas vêm trabalhando no desenvolvimento de padrões para evoluir a WEB, para que, tanto seres humanos quanto computadores, possam entender o conteúdo disponibilizado online. Um dos elementos-chave desta evolução é a construção e manutenção de bases de dados abertas e conectadas [Heath and Bizer 2011]. Nestas bases, as informações são relacionadas entre si para poderem ser consultadas de forma semântica.

Este artigo visa explorar conhecimento estruturado em bases de dados abertas e conectadas a partir da construção de um protótipo de um sistema de perguntas e respostas para auxiliar alunos e professores em suas atividades de estudo da disciplina de geografia.

2. Revisão da literatura

2.1. Sistemas de perguntas e respostas

Sistemas de perguntas e respostas é um tópico de estudo dentro da área de processamento de linguagem natural (PLN), onde o objetivo fundamental é fornecer uma resposta correta e precisa para uma pergunta apresentada em linguagem natural [Cooper and Rüger 2000]. Esses tipos de sistemas podem ser construídos de diferentes maneiras. Uma das maneiras é através de um conjunto de passos que se inicia na seleção e limpeza de um conjunto de documentos servindo de base para consultas. Após, a sentença é quebrada e cada parte é marcada de acordo com sua função gramatical, destacando o foco principal da pergunta. A partir do foco, é possível identificar o tipo de resposta esperado. Por exemplo, quando o foco da pergunta é um lugar, espera-se que o retorno seja uma entidade/conceito de lugar. Depois, uma busca é efetuada na base de documentos para encontrar parágrafos que possam conter palavras-chave relacionadas à pergunta. Obtendo as respostas possíveis, cada uma é avaliada de acordo com critérios do sistema e recebe uma nota. A resposta que obtém a maior nota é apresentada para o usuário. Alguns sistemas podem ter mais ou menos passos dependendo do seu objetivo [Cooper and Rüger 2000].

2.2. Base de dados abertas e conectadas

Bases de dados abertas e conectadas seguem as regras de representação de conhecimento explicitadas em uma ontologia. Os dados armazenados e conectados entre si podem ser facilmente editáveis, pois essas bases podem ser geradas através da extração de informações de portais e plataformas de conteúdo na Internet. A DBPedia, por exemplo, é um projeto que construiu uma base de dados aberta e ligada a partir do conteúdo disponível nas caixas de informações das páginas da Wikipedia [Lehmann et al. 2015]. Além dos dados extraídos, o projeto também relacionou o conteúdo entre si. Para representar essas relações é o utilizado RDF (Resource Description Framework).

RDF é um framework que auxilia na descrição das informações a serem representadas dentro de um conjunto de dados. Ele também descreve todas as relações entre as informações [Heath and Bizer 2011]. Mais especificamente, RDF é um modelo de dados que descreve informações como nodos e arcos dirigidos. Uma informação pode ser representada por uma ou mais tuplas. Uma tupla é composta por três partes: sujeito, predicado/propriedade e objeto. Desta forma, um conjunto de informações em tuplas compõem um grafo RDF [Heath and Bizer 2011].

URI (Uniform Resource Identifier) é um nome usado para identificar e/ou localizar um recurso na Internet. Todos os componentes de uma tupla podem ser um URI. No entanto, um objeto pode ser um valor literal como um número, data, texto, entre outros [Heath and Bizer 2011].

Para consultar base de dados abertas e conectadas modeladas com RDF, foi padronizada uma linguagem de consulta chamada SPARQL pela W3C. SPARQL possibilita a consulta de informações através de um padrão de tuplas onde o sujeito, predicado e objeto da consulta podem ser variáveis [W3C 2008].

3. Trabalhos relacionados

Tendo como base ontologias para gerar perguntas e avaliar respostas, Pereira descreveu o uso de técnicas de processamento de linguagem natural e ontologias aplicados à ambientes virtuais de aprendizagem (AVA) [Pereira and Rigo 2013]. De acordo com o autor, há diversas ferramentas para avaliar textos escritos por estudantes com base em padrões de sentenças. Porém, esses métodos não são suficientes para avaliar textos submetidos à AVAs. Com o objetivo de melhorar a qualidade destas avaliações, Pereira aplicou técnicas de PLN e ontologias para analisar as informações presentes no texto. Para validar sua proposta, Pereira definiu como domínio de conhecimento conceitos de sistemas de gerenciamento de bases de dados (SGBDs) e criou perguntas de forma manual de acordo com o relacionamento de elementos dentro da ontologia. Assim que as informações são extraídas do texto, através de regras linguísticas implementadas por meio de técnicas de PLN, ontologias são consultadas para validar respostas.

Do mesmo modo, Pitrovski propôs um jogo sério para auxiliar alunos durante seus estudos. Baseado no trabalho desenvolvido por Pereira, o autor desenvolveu um quiz sobre os estados do Brasil [Pitrovsk 2015]. As perguntas foram criadas de forma manual pelo autor. Assim que o estudante responde a pergunta, a resposta é analisada com o intuito de extrair componentes-chave para consultar a ontologia do domínio e validar a resposta.

Em 2013, Duma e Klein, também propuseram uma arquitetura para sistemas de geração de linguagem natural através de dados abertos e ligados que aprende, de forma automática, padrões de frases e planejamento de textos a partir de dados RDF, corpus paralelo e documentos textuais. O método proposto por Duma e Klein (2013) extrai *templates* através da mineração de textos em conjunto com a análise de grafos RDF que contêm entidades das sentenças encontradas no processo de mineração.

Os resultados do trabalho proposto pelos autores foram avaliados em comparação com um sistema e um texto escrito por um humano a partir das mesmas informações. O sistema de comparação utilizado gerava sentenças únicas para cada tupla RDF a partir uma análise superficial das palavras do predicado. A arquitetura proposta se mostrou superior ao sistema de frases únicas, mas perdeu em comparação ao texto gerado por humanos.

Perera e Parma, em 2016, criaram um *framework* para transformar uma tupla em uma frase em linguagem natural através da lexicalização. O *framework* é composto por quatro módulos que recebem uma coleção de tuplas previamente anotadas com algumas informações, como, por exemplo, a verbalização da tupla.

O conjunto de tuplas é processado pelos módulos de forma sequencial e caso um módulo encontre um padrão, os demais não são executados. Cada módulo tem uma função que busca por um padrão de lexicalização através de um método específico: metonímia ocupacional, gramática livre de contexto, relacional e, por fim, um conjunto pré-definido de padrões léxicos de tuplas.

Para avaliação do seu trabalho, Perera e Parma (2016) utilizaram regras linguísticas e leitura humana. Para um conjunto de 400 tuplas, o método dos autores gerou 283 padrões precisos de lexicalização e com nível de leitura adequado. Desta forma, os autores concluíram que o *framework* pode servir para geração de padrões de lexicalização

Tabela 1. Algumas das perguntas manualmente relacionadas com propriedades.

Classe	Propriedade	Questão
cidade	dbpedia.org/ontology/demonym	Qual o gentílico desta cidade?
cidade	dbpedia.org/ontology/climate	Qual o clima desta cidade?
cidade	dbpedia.org/ontology/leaderName	Quem é o atual prefeito deste local?
estado	dbpedia.org/ontology/capital	Qual é a capital deste estado?
estado	dbpedia.org/ontology/leaderName	Quem é o atual governador deste estado?
estado	dbpedia.org/ontology/region	Em qual região de seu país este estado pertence?
país	dbpedia.org/ontology/capitalCountry	Qual é a capital deste país?
país	dbpedia.org/ontology/currency	Qual é a moeda utilizada por este país?
país	dbpedia.org/ontology/demonym	Qual o gentílico deste país?

para tuplas de bases de dados abertas e conectadas.

Enquanto os primeiros autores desenvolveram protótipos de sistemas de perguntas e respostas aplicados à educação, os últimos dois trabalhos descritos utilizam de um conjunto de regras e/ou templates para lexicalizar tuplas RDF. Porém, este trabalho de pesquisa se diferencia dos trabalhos citados por utilizar o conhecimento de uma base dados aberta e conectada aplicada à educação.

4. Athena

Athena! foi o nome do protótipo desenvolvido para explorar o conhecimento de bases de dados abertas e ligadas. O projeto DBPedia [Lehmann et al. 2015] foi escolhido como fonte das informações para o sistema. De modo geral, a aplicação desenvolvida permite ao estudante escolher um tópico dentro do domínio de conceitos de geografia regional. Desta forma, estudantes poderiam escolher cidades, estados ou países para responder questões sobre eles.

As questões são geradas de forma manual através de um mapeamento das entidades com suas propriedades. As classes da ontologia que correspondem as entidades de cidade, estado e país possuem propriedades. Para cada propriedade foi elaborada uma pergunta a ser respondida pelo estudante. A Tabela 1 traz algumas relações para as propriedades de cada classe.

Para cada pergunta respondida de forma correta, o estudante ganha um ponto, até finalizar uma sessão de perguntas. Esses pontos alimentam um ranking dentro da aplicação. O ranking da aplicação contém um conjunto de níveis, cada qual é atingido a partir do número de acertos do estudante. O primeiro nível é intitulado de ‘Estudante’ enquanto o último nível tem a titulação de ‘Doutor’. Para participar do ranking, o estudante precisa utilizar suas credenciais do Facebook. Desta forma, o sistema consegue, a partir dos dados de identificação do estudante, guardar as informações relativas às sessões de perguntas e respostas do usuário. Além disso, o sistema também permite o compartilhamento do resultado de cada sessão e as mudanças de níveis.

4.1. Jornada do usuário

A jornada do usuário dentro do sistema começa em uma tela que oferece duas opções para iniciar a aplicação. O usuário pode escolher identificar-se no sistema com suas credenciais do Facebook ou acessá-lo de forma anônima (Figura 1).

Figura 1. Tela inicial do sistema.

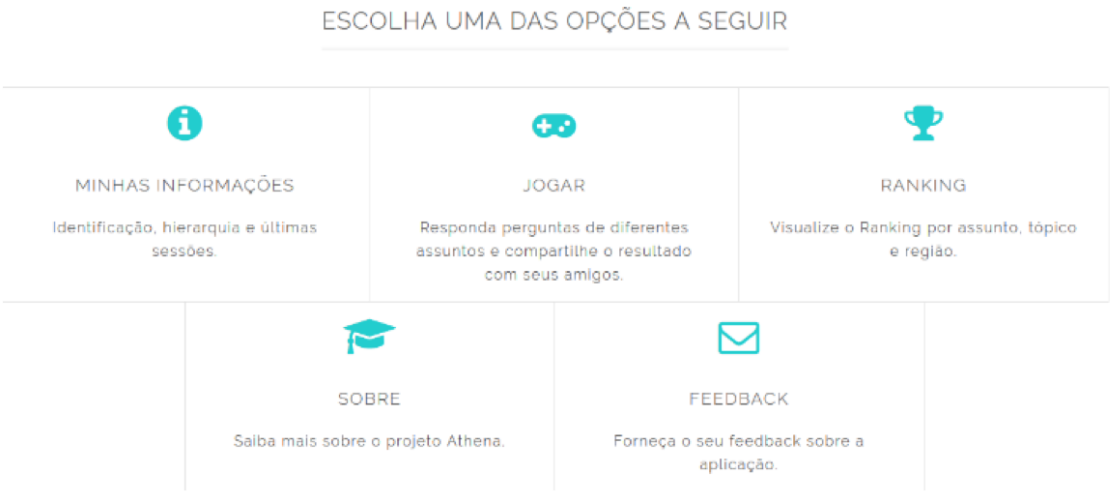

Figura 2. Tela do painel do sistema após identificação.

A identificação do usuário possibilita o rastreamento de suas sessões para composição do ranking. Após a identificação do usuário, são oferecidas algumas opções para uso da aplicação, conforme mostra a Figura 2. Dentre as opções estão a visualização das informações das sessões do usuário, iniciar uma nova sessão do jogo, visualizar o ranking, fornecer feedback sobre a aplicação e visualizar as informações do projeto do protótipo.

A principal funcionalidade é a geração de perguntas e validação das respostas. Ao acessar esse recurso pela opção Jogar, o usuário precisa selecionar um tópico (Figura 3) para que uma sessão de perguntas se inicie.

As perguntas são apresentadas com um campo de texto para que o usuário informe a resposta que achar conveniente para o questionamento. Além disso, outras opções são

Figura 3. Tela para escolha do tópico para iniciar a geração de perguntas.

apresentadas para que o usuário possa ver a resposta correta, desista, ou ouça a pergunta em linguagem natural, conforme mostra a Figura 4.

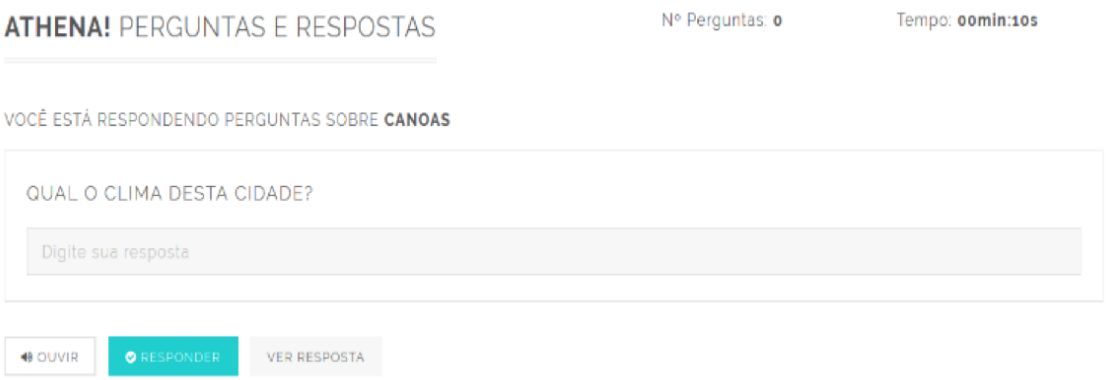

Figura 4. Tela da sessão de perguntas e respostas contendo a opção para receber a resposta do usuário.

Caso o usuário acerte as perguntas, novos questionamentos são gerados. Se todos os questionamentos mapeados para o tópico escolhido forem respondidos corretamente pelo usuário, o sistema possibilita a escolha de um novo tópico dentro da mesma sessão. No entanto, caso um usuário erre a resposta de pergunta, a aplicação oferece a resposta correta ao questionamento e uma opção para iniciar uma nova sessão de perguntas, conforme Figura 5.

4.2. Arquitetura

Para permitir a construção dessas funcionalidades, o sistema foi desenhado para conter quatro componentes, ilustrados na Figura 6.

A camada de *front-end* da aplicação é a interface do usuário. A interface foi desenvolvida como uma aplicação WEB utilizando o AngularJS baseado em um template

Figura 5. Tela após o encerramento de uma sessão.

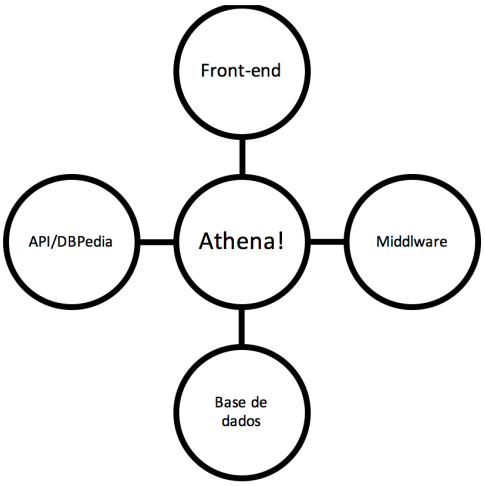

Figura 6. Arquitetura do Athena!.

CSS com código aberto disponível na Internet [Templated 2015].

A camada de front-end se comunica com um *middleware* para acessar serviços. Esta camada foi desenvolvida utilizando NodeJS. Os serviços expostos pelo *middleware* se conectam com outros provedores consultando dados, fazendo um processamento destes dados e retornando informações para o front-end. O principal provedor externo à aplicação é o projeto DBPedia. Uma cópia local do dump(2016/04) do capítulo em português do projeto foi instalado em um servidor na nuvem para ser utilizado pelo Athena!. Outro provedor amplamente utilizado é a interface de programação do Facebook para autenticação e compartilhamento de conteúdo. Além disso, o *middleware* também é responsável pela gestão dos dados a serem gravados no banco de dados.

Dentro do *middleware* há uma estrutura modular, onde cada módulo é responsável por um passo no ciclo de vida do sistema de perguntas e respostas. Mais especificamente, seguindo o fluxo de uso do sistema, os módulos são os seguintes:

1. Pesquisa do tópico do usuário na DBPedia – uma pesquisa por tópicos que combinam com a palavra-chave digitada pelo usuário é feita utilizando SPARQL na DBPedia. O retorno desta consulta retorna opções que o usuário pode iniciar uma nova sessão do jogo.
2. Busca pelas perguntas mapeadas para o conceito escolhido – uma busca no mape-

amento de perguntas e conceitos é executada.
3. Pesquisa das propriedades disponíveis no conceito escolhido – uma consulta SPARQL na DBPedia verifica se a instância do conceito escolhido pelo usuário contém a propriedade mapeada. Caso contrário, é informado ao usuário que o tópico não está disponível para questionamentos.
4. Apresentação da pergunta – encontrando uma propriedade válida para a instância do conceito escolhido pelo usuário, a pergunta é apresentada para ele.
5. Busca pela resposta da pergunta – após o usuário informar a resposta para a pergunta, um pré-processamento ocorre na resposta. Com base na pergunta apresentada, uma consulta é executada na DBPedia para buscar a resposta da pergunta.
6. Validação da resposta informada pelo usuário – após receber o resultado da consulta pela resposta na DBPedia, o mesmo pré-processamento ocorre para tratar carácteres inválidos. Uma comparação é feita entre a resposta informada e a informação que retornou da consulta da base de dados.

Além disso, para permitir o rastreamento das sessões dos estudantes e manter o ranking, uma base de dados utilizando MongoDB foi acoplada à arquitetura. Nesta base de dados, todas a sessões são registradas contendo o número de perguntas respondidas corretamente, tempo total de sessão e número de perguntas apresentadas.

5. Avaliações e resultados

Por meio de amostragem por conveniência, turmas de ensino fundamental e médio de algumas escolas foram convidadas a utilizar o sistema e preencher uma rápida pesquisa. Este material tinha como objetivo indicar o potencial do uso de base de dados abertas e ligadas para construção de aplicações a serem utilizadas no contexto educacional por estudantes e professores.

Um total de 10 professores e 61 alunos de duas escolas responderam a pesquisa após interagirem com o sistema. As escolas que participaram da pesquisa situam-se no munícipio de Canoas dentro da região metropolitana de Porto Alegre, no Rio Grande do Sul.

Na primeira escola, que é administrada pelo governo estadual, 4 turmas participaram das pesquisas. Duas turmas de ensino fundamental, 7º e 9º ano, e outras duas turmas de ensino médio, 2º e 3º anos. Por questões de disponibilidade, as interações por parte dos alunos com o sistema foram executadas nos períodos da tarde e noite, o que aumentou a variação de idade dos estudantes. Estudantes de 12 a 25 anos tiveram a oportunidade de utilizar a aplicação. Na segunda escola, administrada pelo governo municipal, somente uma turma de 9º ano do ensino fundamental teve autorização para participar da pesquisa.

Para as duas escolas, o mecanismo de apresentação do sistema foi o mesmo. Nos primeiros 5 minutos foram explorados os motivadores do desenvolvimento do protótipo, seguidos de uma curta demonstração do sistema. A partir disto, o uso do sistema foi liberado para os estudantes com a provocação de tentar acertar o maior número de perguntas. Após 10 minutos de uso do sistema, os participantes foram convidados a responder um questionário.

Para os professores das duas escolas, o procedimento foi semelhante. Uma apresentação do protótipo e os motivadores do desenvolvimento foram introduzidos, seguidos de uma curta demonstração, para que os professores interagissem com a aplicação.

Tabela 2. Perguntas manualmente relacionadas com propriedades.

#	Questão
1	As questões eram fáceis?
2	A experiência de usabilidade com a aplicação foi satisfatória?
3	A aplicação pode apoiar atividades dentro de sala de aula?
4	A aplicação pode apoiar atividades fora de sala de aula?
5	A aplicação pode ser usada pelos alunos de ensino fundamental?
6	A aplicação pode ser usada pelos alunos de ensino médio?

Tabela 3. Resultado das perguntas para professores.

Perg.	Disc. plen.	Discordo	Nem disc./conc.	Concordo	Conc. plen.
1	1	1	0	4	4
2	0	1	0	0	9
3	0	0	0	1	9
4	0	0	0	0	10
5	0	0	0	1	9
6	0	0	0	1	9

Para os dois grupos, professores e alunos, o mesmo conjunto de perguntas, seguindo o modelo de escala LIKERT, foram apresentadas através de um sistema online. O questionário está disposto na Tabela 2. Os números obtidos na avaliação encontram-se nas Tabelas 3 e 4 para professores e alunos, respectivamente.

A partir das questões apresentadas aos dois grupos, todos os professores que participaram da avaliação do protótipo indicaram que a exploração de conhecimento de bases de dados abertas e ligadas pode ser benéfico para uso em aplicações educacionais, tanto dentro quanto fora de sala de aula. Além disso, os professores também em sua maioria concordaram com o uso da aplicação para alunos de ensino fundamental e médio, mas destacaram que algumas perguntas geram dificuldades maiores para determinados níveis de ensino. Por exemplo, uma das questões mapeadas trata-se do nome do prefeito da cidade, que poucos alunos do ensino fundamental têm conhecimento.

Para os alunos, a maioria concordou com o fato que o protótipo e as tecnologias envolvidas na sua construção podem ser benéficas para uso por alunos e professores como recurso didático dentro e fora de sala de aula. Similarmente aos professores, alguns alunos acharam questionamentos difíceis para seus níveis de ensino.

Tabela 4. Resultado das perguntas para alunos.

Perg.	Disc. plen.	Discordo	Nem disc./conc.	Concordo	Conc. plen.
1	3	4	10	24	20
2	1	2	3	16	39
3	1	0	0	14	46
4	0	2	3	13	43
5	0	3	5	20	33
6	1	0	5	17	38

6. Conclusão

As informações publicadas na Internet, quando estruturadas, podem servir de base de conhecimento para sistemas especialistas com aplicações nas mais diversas áreas. Este artigo descreve de forma resumida o trabalho desenvolvido em torno de um protótipo de um sistema de perguntas e respostas que explora o conhecimento disponível em bases de dados abertas e conectadas.

Após a construção do protótipo, um grupo de alunos e professores foram convidados, por meio de uma amostragem por conveniência, para compartilharem suas opiniões a respeito do sistema desenvolvido. Tanto alunos quanto professores, concordaram em indicar a aplicação desenvolvida e suas tecnologias como benéficas para uso dentro e fora de sala de aula.

Portanto, este trabalho desenhou e desenvolveu um protótipo de sistemas de perguntas e respostas que exploram o conhecimento de bases de dados abertas e conectadas. O software desenvolvido, por ter uma arquitetura modular, favorece a reutilização para trabalhos futuros aprimorarem cada um dos módulos. Além disso, esta primeira avaliação conduzida em conjunto com alunos e professores indica que trabalhos futuros podem explorar com mais precisão o potencial de sistemas de perguntas e respostas apoiados em bases de dados abertas e ligadas.

Outros trabalhos futuros também podem ser realizados no módulo de geração de linguagem natural. Também é possível estudar e desenvolver meios de gerar linguagem natural a partir do relacionamento entre o conhecimento das bases de dados de forma automática. No módulo de validação também há a possibilidade de consultar léxicos para encontrar sinônimos das respostas para aumentar o nível de assertividade das perguntas.

Referências

Cooper, R. J. and Rüger, S. M. (2000). A simple question answering system. In *TREC*.

Heath, T. and Bizer, C. (2011). Synthesis lectures on the semantic web: theory and technology. *Linked Data: Evolving the Web into a Global Data Space*, 1:1–136.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Mesinovic, A. (2013). How to link marketing and sales. Disponível em $http : //www.business2community.com/content - marketing/how - to - link - content - marketing - and - sales - 0475109$, acesso em 01/05/2017.

Pereira, F. R. and Rigo, S. J. (2013). Utilização de processamento de linguagem natural e ontologias na análise qualitativa de frases curtas. *RENOTE*, 11(3).

Pitrovsk, R. (2015). Um jogo educacional com o uso de processamento de linguagem natural e ontologias. Monografia (Bacharel em Ciência da Computação), UNISINOS (Universidade do Vale do Rio dos Sinos), São Leopoldo, Brasil.

Templated (2015). Typify. Disponível em $https : //templated.co/typify$, acesso em 16/04/2016.

W3C (2008). Sparql query language for rdf. Disponível em $https : //www.w3.org/TR/rdf - sparql - query.$, acesso em 20/04/2017.