Uma Abordagem Baseada em Dados Abertos Conectados e *Chatbot* para Disponibilizar o Catálogo de Cursos da Rede Federal de Educação Profissional, Científica e Tecnológica

Antônio J. Moraes Neto[1], Clênio E. Silva[2], William F. Anjos[2], Fabiano A. Dorça[2]

[1]Instituto Federal de Brasília (IFB) – Brasília, DF – Brasil

[2]Universidade Federal de Uberlândia (UFU) – Uberlândia, MG – Brasil

antonio.neto@ifb.edu.br,{clenio.silva, williamferreira, fabianodor}@ufu.br

Abstract. *In Brazil there are 661 federal campuses in operation, offering courses of Professional and Technological Education. Information about these courses is disseminated on the websites of more than sixty institutions that form a network of Federal Institutes of Education, Science and Technology. Ontology engineering allows us to produce an effective structure for the publication of connected open data in order to contribute to the availability, quality and reliability of data about such courses. As a result, are proposed an ontology to catalog the courses offered by these educational institutions and a chatbot to facilitate access to this information in a web environment.*

Resumo. *Na Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) há 661 campi em funcionamento, oferecendo cursos de Educação Profissional e Tecnológica (EPT), cujas informações estão difusas nos sites das mais de sessenta instituições de ensino da RFEPCT. A engenharia de ontologia permite produzir uma estrutura eficaz para a publicação de dados abertos conectados a fim de contribuir para a disponibilidade, qualidade e confiabilidade dos dados acerca de tais cursos. Como resultado, são propostos uma ontologia para catalogar cursos de EPT ofertados na RFEPCT e um chatbot para facilitar o acesso a tais informações em ambiente web.*

1. Introdução

Atualmente a Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) conta com 661 *campi* em funcionamento que ofertam cursos, em diversos níveis e modalidades educacionais, a fim de promover a formação profissional de pessoas com perfis diversificados para atuarem em variadas áreas do mundo do trabalho. Essa é uma situação emblemática de como as informações são geradas e disponibilizadas na Internet sem que estejam estruturadas, de modo a facilitar sua compreensão por parte dos interessados. Buscando modificar realidade como esta, esforços têm sido realizados a fim de "disponibilizar dados e produzir tecnologias web que permitam criar um ecossistema de produção e consumo de dados com o objetivo de agilizar a descoberta de novos conhecimentos e agregar valor a qualquer informação disponibilizada livremente na Internet" [Isotani e Bittencourt 2015 p. 17].

Os dados dos cursos de Educação Profissional e Tecnológica estão disponíveis nos sites das sessenta e duas instituições de ensino integrantes da RFEPCT [Brasil

DOI: 10.5753/cbie.sbie.2020.1263

2020]. Contudo não há estruturação destes sites na perspectiva de dados abertos conectados, viabilizando que estejam disponíveis e acessíveis, em formato conveniente e em lugar indexado, a fim de que possam ser reusados e redistribuídos por qualquer pessoa e máquina, com permissão para que isso ocorra sem restrições. A partir dessa lacuna este trabalho de pesquisa tem o objetivo de desenvolver o Catálogo de Cursos da RFEPCT, em forma de dados abertos conectados, e propor um *chatbot* que facilite o acesso a este, contribuindo assim para a disponibilidade, qualidade e confiabilidade dos dados acerca desses cursos.

2. Fundamentos Bibliográficos

Uma solução que vem sendo adotada com o objetivo de melhorar o atendimento a clientes, proporcionando soluções imediatas por parte de empresas é o uso de agente conversacional, ou *chatbot,* como ferramenta para oferecer melhor experiência de atendimento ao cliente [Grazioli 2017]. Este agente virtual é capaz de interagir com uma pessoa por meio de texto ou áudio, respondendo a um grande volume de solicitações com qualidade similar aos atendentes humanos. Isso é possível devido ao uso da Inteligência Artificial (IA), fornecendo um conjunto de técnicas que permitem ao *chatbot* aprender com cada conversa de maneira a melhorar as próximas interações.

Pesquisadores vêm estudando a aplicação de *chatbots* em contexto educacional como apoiadores de tutoria e ensino. Um exemplo é apresentado por Lucchesi et al. (2018) que propõem a Metis (Mediadora de Educação em Tecnologia Informática e Socializadora), projetada para conversar com estudantes por meio de uma interface de chat, disponibilizada pela Internet, podendo fornecer, em qualquer momento, informações educacionais e apontar para fontes externas de dados.

Ampliando as possibilidades de aplicação da IA na Educação (IAED), podemos considerar ainda "algumas iniciativas que visam estimular a produção e uso de dados abertos conectados na educação", como a LinkedEducation.org, "uma plataforma aberta que visa promover o uso de dados conectados educacionais mediante o incentivo ao compartilhamento de recursos e informações relacionadas à educação" [Bandeira et al. 2015 p. 65]. Nesta são divulgadas boas práticas e potenciais relações entre recursos a fim de contribuir para a interligação eficiente de dados educacionais na web.

Com base no contexto exposto, neste trabalho é descrito o desenvolvimento de um Catálogo de Cursos da RFEPCT e de um *chatbot* para facilitar o acesso de interessados a informações desses cursos. Busca-se beneficiar a divulgação destes, por meio de uma interface de chat que permite fazer perguntas em linguagem natural e obter respostas a partir dos dados contidos em uma ontologia proposta, que é a base de conhecimento do *chatbot*. São diversos os interessados, podendo ser tanto candidatos a cursos e representantes de organizações sociais que atuam em áreas correlatas, quanto integrantes das comunidades acadêmicas que terão mais disponibilidade de informações acerca dos cursos ofertados pela RFEPCT no Brasil. Pesquisas em Educação Profissional e Tecnológica (EPT) contarão com referencial para ampliar o conhecimento disponível nesta aplicação da IAED, podendo ainda contribuir para sua melhoria.

3. Metodologia

Com objetivo de construir uma ontologia para estruturação do conhecimento sobre o Catálogo de Cursos da RFEPCT foi adotada a metodologia *Ontology Development 101*, definida por Noy e Mcguinness (2001), que contém sete etapas a serem realizadas para o desenvolvimento de uma ontologia, ilustradas na Fase 1 do fluxograma que consta na Figura 1. A seguir a engenharia da ontologia proposta neste trabalho está descrita a partir da realização dessas sete etapas, tendo sido desenvolvida com o apoio do software livre para edição de ontologias Protégé[1].

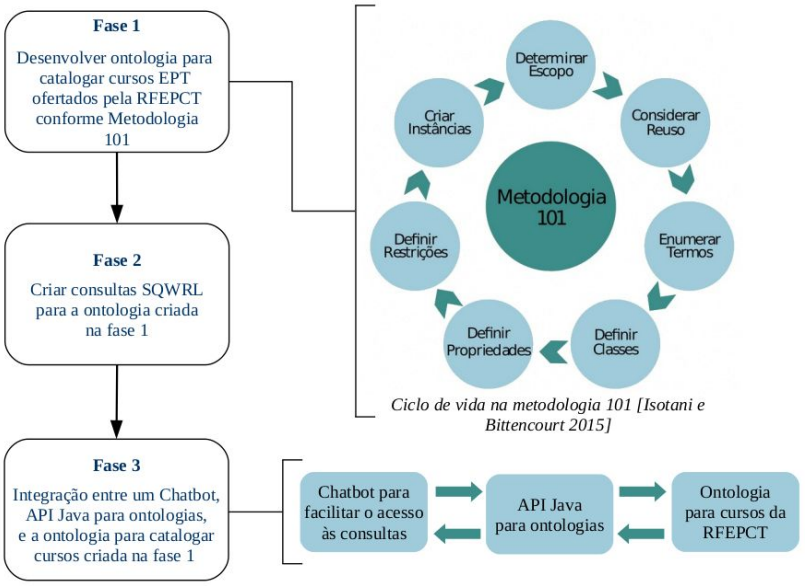

Figura 1. Fluxograma da Metodologia Adotada [Autoria própria]

4. Engenharia da Ontologia Catálogo de Cursos da RFEPCT

Como desenvolver um catálogo de cursos ofertados pela RFEPCT, em forma de dados abertos conectados, a fim de contribuir para ampliar a disponibilidade, qualidade e confiabilidade dos dados acerca desses cursos? Em função de responder a essa questão de pesquisa os autores construíram, de acordo com as etapas de Noy e Mcguinness (2001), a ontologia Catálogo de Cursos da RFEPCT.

4.1. Determinar Escopo

O objetivo da ontologia é catalogar os cursos ofertados pela RFEPCT, podendo ser usada para auxiliar a tomada de decisão por parte de:

- Professores, quanto ao desenvolvimento e aprimoramento de recursos educacionais como planos de aulas e Projeto Pedagógico de Curso (PPC);
- Gestores escolares, a fim de elaborarem o planejamento educacional, podendo constituir uma referência às instituições de ensino;
- Outros interessados, como possíveis estudantes, empregadores e organizações parceiras, a partir da caracterização de cursos ofertados pela RFEPCT.

[1] Protégé: https://protege.stanford.edu/.

4.2. Considerar Reuso

Foi considerado o reuso do vocabulário Schema.org, permitindo interoperabilidade com a solução proposta. No segundo nível da hierarquia deste existem tipos especializados, cada um com seu próprio vocabulário, dentre eles o *Educational Organization*[2]. Desta forma, os autores consideraram inadequado o reuso de vocabulário na ontologia proposta.

4.3. Enumerar Termos

Para enumerar os termos que compõem a ontologia Catálogo de Cursos da RFEPCT é relevante observar a estrutura da educação formal brasileira que, de acordo com Distrito Federal (2018), é desenvolvida em:

- Níveis de educação, quais sejam: Educação Básica e Educação Superior;
- Etapas da Educação Básica, que são Educação Infantil, Ensino Fundamental e Ensino Médio;
- Modalidades da educação, sendo: Educação de Jovens e Adultos (EJA); Educação Especial; Educação Profissional e Tecnológica (EPT); Educação do Campo; Educação Indígena; Educação Quilombola e Educação a Distância (EAD).

Neste cenário, a EPT é um processo desenvolvido em articulação com a Educação Básica, nos Ensinos Fundamental e Médio, e Educação Superior, em suas diferentes modalidades. No Brasil é desenvolvida por cursos e programas de:

- Formação inicial e continuada (FIC);
- Educação Profissional Técnica de Nível Médio, podendo ser ofertada nas formas: Concomitante; Integrada e Subsequente;
- Educação Profissional Tecnológica, de Graduação e Pós-graduação.

Outros conceitos considerados contidos na ontologia proposta são: Áreas de atuação com base nos eixos tecnológicos do Catálogo Nacional de Cursos Técnicos (CNCT) e nas Áreas de Formação do Painel Lattes; Certificações e habilitações a partir da Classificação Brasileira de Ocupações (CBO); Instituições de ensino ofertantes de cursos da EPT e Localidades por UF.

Quanto a questões de competência da ontologia Catálogo de Cursos da RFEPCT, espera-se com seu uso poder responder às seguintes questões:

- Quais cursos estão disponíveis para determinada área de atuação?
- Que cursos são ofertados em determinada modalidade de ensino, como a EAD?
- Em uma UF quais os cursos de determinado nível educacional, por exemplo licenciatura, estão sendo ofertados?
- Tendo encontrado um curso de interesse, onde encontrar as informações detalhadas, como o seu PPC?

[2] Educational Organization: https://schema.org/EducationalOrganization.

4.4. Definir Classes

A partir do que foi determinado nos itens anteriores os autores definiram a estrutura hierárquica de classes, com base na taxonomia dos termos levantados, apresentada na Figura 2. Os demais termos relevantes determinados, que não constam nesta figura, foram definidos como instâncias das respectivas classes.

Figura 2. Estrutura Hierárquica de Classes [Autoria própria]

4.5. Definir Propriedades

As propriedades das classes, em inglês *slots*, que representam os relacionamentos de algumas classes com outras, são: Campus_e_polo_de; E_area_de_atuacao_de; E_cidade_de; E_localidade_de; E_modalidade_de; E_municipio_de; E_nivel_educacional_de; E_ofertado_por; E_ofertante_de; E_titulacao_de; Esta_localizado_em; Oferece_titulacao_em; Pertence_a_instituicao; Polo_e_polo_de; Tem_area_de_atuacao; Tem_modalidade; Tem_nivel_educacional.

As propriedades de dados representam atributos a serem cadastrados na etapa final, de instanciamento, sendo: Area_tem_nome; Campus_tem_link; Campus_tem_nome; Cidade_tem_nome; Curso_tem_link; Curso_tem_nome; Instituicao_tem_link; Instituicao_tem_nome; Instituicao_tem_sigla; Municipio_tem_nome; Polo_tem_nome; Titulo_tem_nome; Uf_tem_nome; Uf_tem_sigla.

4.6. Definir Restrições

Nesta etapa é necessário definir restrições para as propriedades tanto de classes quanto de dados. Na Tabela 1 constam aquelas para as primeiras.

Tabela 1. Definição de Restrições: Propriedades das Classes [Autoria própria]

Slot	***Domain***	***Range***	***InverseOf***
E_ofertado_por	Curso	Campus	E_ofertante_de
E_ofertante_de	Campus	Curso	E_ofertado_por

Tem_nivel_educacional	Curso	Nivel_Educacional	E_nivel_educacional_de
E_nivel_educacional_de	Nivel_Educacional	Curso	Tem_nivel_educacional
Tem_modalidade	Curso	Modalidade	E_modalidade_de
E_modalidade_de	Modalidade	Curso	Tem_modalidade
Pertence_a_instituicao	Campus	Instituicao_de_ensino	
Campus_e_polo_de	Campus	Instituicao_de_ensino	
Polo_e_polo_de	Polo	Campus	
Tem_area_de_atuacao	Curso	Area_de_atuacao	E_area_de_atuacao_de
E_area_de_atuacao_de	Area_de_atuacao	Curso	Tem_area_de_atuacao
Oferece_titulacao_em	Curso	Titulo	E_titulacao_de
E_titulacao_de	Titulo	Curso	Oferece_titulacao_em
Esta_localizado_em	Campus	Localidade	E_localidade_de
E_localidade_de	Localidade	Campus	Esta_localizado_em
E_cidade_de	Cidade	Uf	
E_municipio_de	Municipio	Uf	

As restrições para as propriedades de dados foram especificadas quanto a:

- *Domain*, que define a propriedade de classe à qual a restrição se aplica;
- *Ranges*, que determina o tipo de dado aceito, como *number*, *string*, *anyURI* para link de Internet;
- *Data range expression*, podendo conter uma lista de termos possíveis, como no caso das propriedades Uf_tem_nome e Uf_tem_sigla;
- *Cardinality*, para, quando for o caso, apontar a quantidade mínima ou máxima necessária em cada instanciamento.

4.7 Criar Instâncias

As instâncias foram criadas a partir dos cursos ofertados por alguns *campi* de instituições da RFEPCT, com diversidade de níveis, etapas e modalidades da educação. Primeiramente, foram inseridos dados das classes elementares para os cursos como Area_de_atuacao, Instituicao_de_ensino, Campus, Localidade, Titulo. Posteriormente, juntamente com os cursos, foram cadastrados os dados referentes às classes Nivel_educacional e Modalidade_de_ensino. Vários ajustes na ontologia proposta foram feitos nesta etapa. Dessa forma foi possível avançar para as Etapas 2 e 3 da metodologia adotada neste trabalho.

5. Criação de Consultas

Após a construção da ontologia, o próximo passo foi a criação de consultas para extrair informações que respondessem as questões inerentes à competência da ontologia, definidas inicialmente na enumeração de termos. Para tanto, foi utilizada a *Semantic Query-Enhanced Web Rule Language* (SQWRL), linguagem de consultas que fornece operadores semelhantes aos da *Structured Query Language (*SQL) para extrair informações de ontologias desenvolvidas em *Ontology Web Language* (OWL). Segundo O'connor (2016), a linguagem SQWRL é baseada em *Semantic Web Rule Language* (SWRL), que por sua vez, é uma linguagem de regras da web semântica, na qual basicamente as regras propostas são em forma de uma implicação entre um antecedente e consequente. O significado pretendido pode ser lido como: sempre que as condições especificadas no antecedente são mantidas, as condições especificadas no consequente também devem ser mantidas.

A SQWRL, utilizada para criar consultas na ontologia proposta, usa um antecedente de regra SWRL como uma especificação de padrão e substitui a regra consequente pelos operadores de seleção SQWRL. Nesta proposta as consultas foram criadas para responder às questões inerentes à competência da ontologia proposta, de acordo com as classes, propriedades de objetos e de dados criadas na ontologia.

6. O *Chatbot* para Facilitar o Acesso ao Catálogo de Cursos da RFEPCT

Para desenvolvimento do *chatbot* foram consideradas técnicas de dois subcampos da IA, sendo Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM). Os dois subtópicos a seguir abordam e detalham o uso dessas duas técnicas para o desenvolvimento do agente conversacional.

6.1. Processamento de Linguagem Natural

PLN estuda a capacidade e as limitações de uma máquina em entender a linguagem dos seres humanos [Rodrigues 2017]. Para modelagem linguística, possibilitando que a máquina entenda a linguagem natural, são necessários pré-processamentos que extraem a estrutura da língua com informações relevantes, reduzindo o vocabulário, facilitando o trabalho com dados não estruturados e possibilitando o processamento computacional. O algoritmo de pré-processamento foi desenvolvido seguindo as etapas de pré-processamento descritas em Bird, Klein e Loper (2009), as quais são normalização, remoção de palavras mais frequentes (*stopwords*), remoção de numerais e redução de palavras flexionadas a seus radicais (*stemming*).

A normalização faz tratamentos no texto como a ‘tokenização’, passagem de letras maiúsculas para minúsculas, remoção de caracteres especiais, remoção de *tags* HTML, links HTTP, dentre outros. Além do pré-processamento descrito foram aplicadas outras duas técnicas, sendo *n-gram* com aplicação de *bigrams* para criação do vocabulário e do *bag-of-words* para criação dos vetores de entrada para o algoritmo de aprendizado de máquina. O pré-processamento é aplicado em todas as frases da base de dados para treinamento do modelo de aprendizagem e na fase de teste.

6.2. Aprendizado de Máquina

AM é atualmente muito usado para análise de dados, que aplica modelos matemáticos proporcionando que esses aprendam a identificar padrões e tomar decisões com base nos dados previamente conhecidos, ocorrendo o treinamento [Géron 2017]. Para esse trabalho foi adotado como modelo de aprendizagem o teorema de *Bayes*, um algoritmo clássico para classificação em PLN. No desenvolvimento do algoritmo proposto para a classificação dos textos foi usada a biblioteca de AM *Scikit-learn*. Foram desenvolvidas duas aplicações *RESTful*, uma com a linguagem Java e outra com a linguagem Python. A aplicação Java faz uso da biblioteca SWRLAPI que permite a realização de consultas na ontologia desenvolvida. Já a aplicação Python é responsável pelo PLN, treinamento do modelo de aprendizagem e classificação dos textos de entrada. Na Figura 3 consta o diagrama com o fluxo de comunicação da aplicação.

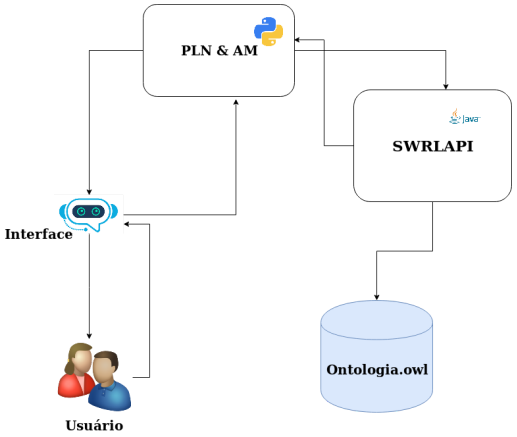

Figura 3. Fluxo de Comunicação da Aplicação Desenvolvida [Autoria própria]

Após a conclusão das fases 1 e 2 da metodologia adotada, foram definidas as consultas SQWRL na aplicação Java a fim de realizá-las na ontologia. Para cada consulta foi gerado um rótulo, sendo 18 com rótulos “q1”, “q2”, “q3” … “q18”. O modelo de aprendizagem da aplicação Python realiza a predição para um texto de entrada dentre 18 classes, sendo elas “q1”, “q2”, “q3” … “q18”, depois faz uma requisição na aplicação Java, passando a predição encontrada. A aplicação Java devolve o resultado da consulta SQWRL realizada na ontologia para a aplicação Python, que processa e retorna os resultados para a interface do *chatbot*. Ao final a interface apresenta em uma tela de chat o resultado para o usuário.

7. Resultados

Para validação do *chatbot* foram propostas algumas perguntas que avaliam a acurácia do mesmo em relação às competências da ontologia Catálogo de Cursos da RFEPCT. Para o algoritmo de classificação foi considerado um limite para a predição das classes, que, se corresponder abaixo de 60%, a classe predita é indefinida, garantindo assim que o *chatbot* tenha maior assertividade nas respostas.

O experimento foi conduzido pelos autores do trabalho com a realização de perguntas em linguagem natural ao *chatbot*. Em prol da generalização do modelo de PLN foram criadas 3 perguntas diferentes para cada uma das 18 consultas SQWRL, totalizando 54 questionamentos. Na Figura 4 está uma sequência de perguntas feitas

pelo usuário, em caixas de diálogo verdes, com as respostas do *chatbot*, nas caixas brancas.

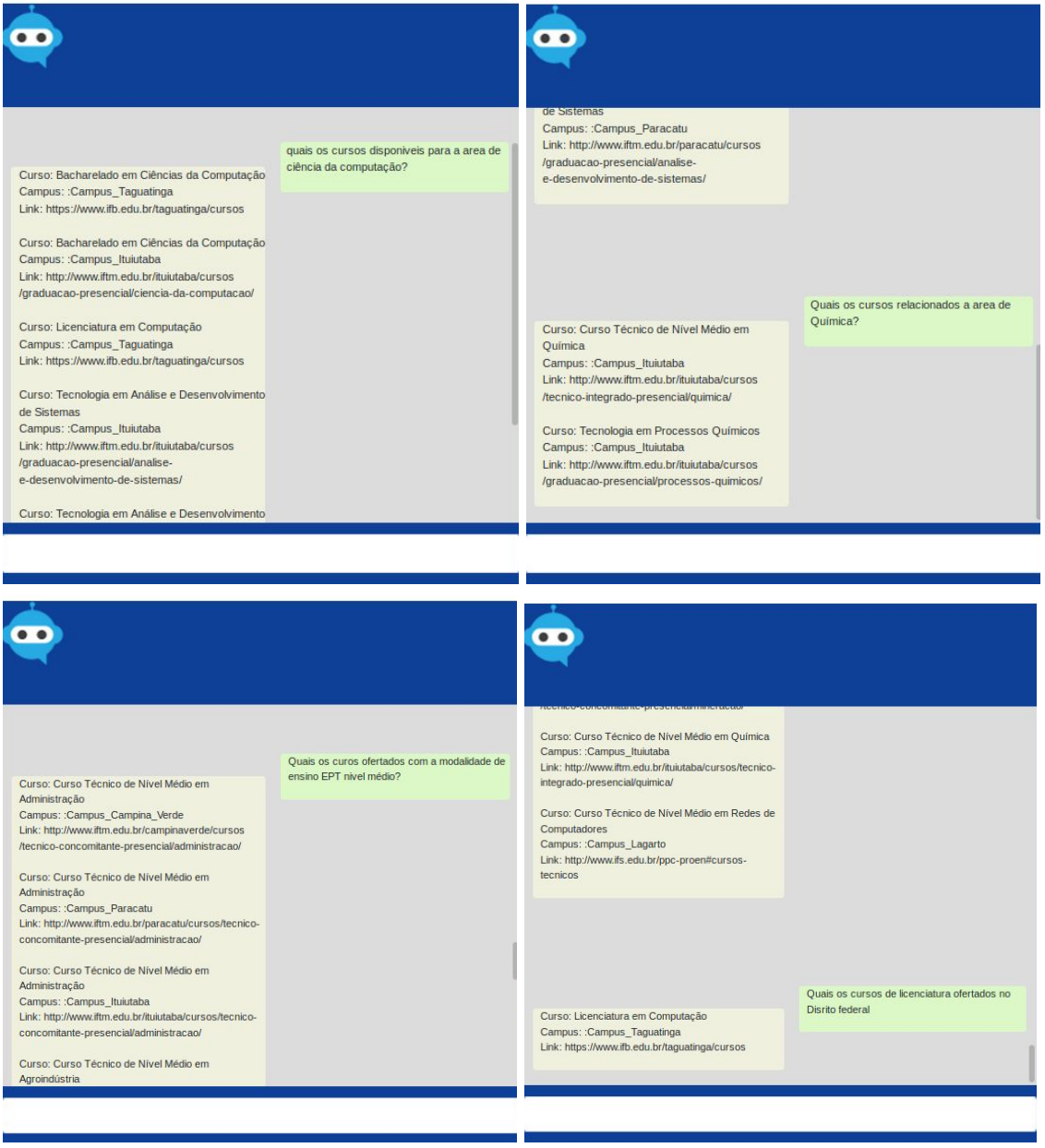

Figura 4. Interações de Usuário com o *Chatbot* Desenvolvido [Autoria própria]

Com os resultados apresentados foi possível validar esta abordagem, sendo que, para as perguntas propostas o *chatbot* conseguiu responder todas, o que demonstra a eficácia da arquitetura definida na ontologia Catálogo de Cursos da RFEPCT, incluindo as inferências para a associação de informações a fim de produzir conhecimento acerca dos cursos ofertados pela RFEPCT.

8. Conclusão e Trabalhos Futuros

Neste artigo foram propostos uma ontologia para catalogar cursos de Educação Profissional e Tecnológica e um *chatbot* para facilitar o acesso dos interessados a informações acerca de cursos da RFEPCT, rede que conta com centenas de *campi* em

funcionamento no Brasil. Por meio da base de conhecimento implementada pôde-se extrair informações relevantes tanto para estudantes, professores e administradores acadêmicos, quanto para outras organizações e candidatos interessados nesses cursos, trazendo diversas vantagens em relação a acessibilidade destas informações.

Como trabalho futuro pretende-se implementar ou integrar a ontologia Catálogo de Cursos da RFEPCT e o *chatbot* desenvolvidos em alguma plataforma web a fim de contribuir para a disponibilidade, qualidade e confiabilidade dos dados acerca de tais cursos e para o crescimento da web semântica, disponibilizando dados educacionais públicos de forma estruturada e aberta, o que ampliará o acesso, reuso e redistribuição desses dados por parte de interessados, seja pessoa, grupo ou máquina.

Referências

Bandeira, J., Ávila, T., Alcantara, W., Barbosa, A., Bittencourt, I., Isotani, S. (2015) "Dados abertos conectados para a Educação", http://www.br-ie.org/pub/index.php/pie/article/view/3551.

Bird, S., Klein, E., Loper, E. (2009) "Natural Language Processing With Python: Analyzing Text with the Natural Language Toolkit", O'REILLY, ISBN:978-0-596-51649-9.

Brasil, RFEPCT. (2020) "Instituições da Rede Federal", http://portal.mec.gov.br/rede-federal-inicial/instituicoes.

Distrito Federal, Conselho de Educação do Distrito Federal. (2018) "Resolução Nº 1/2018", http://cedf.se.df.gov.br/resolucoes/resolicao-cedf.

Grazioli, D. (2017) "Tendências: por que usar chatbots no relacionamento com o cliente",http://agenciamulticom.com.br/site/tendencias-por-que-usar-chatbots-no-relacionamento-com-o-cliente/.

Géron, A. (2017) "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems", O'reilly, ISBN: 1491962291.

Isotani, S., Bittencourt, I. (2015) "Dados Abertos Conectados", https://ceweb.br/publicacao/livro-dados-abertos/.

Lucchesi, I., Silva, A., Abreu, C., Tarouco, L. (2018) "Avaliação de um Chatbot no Contexto Educacional: um Relato de Experiência com Metis", https://doi.org/10.22456/1679-1916.85903.

Noy, N., Mcguinness, D. (2001) "Ontology Development 101: A Guide to Creating Your First Ontology", https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

O'connor, M. (2016) "SQWRL: Semantic Query-Enhanced Web Rule Language", https://github.com/protegeproject/swrlapi/wiki/SQWRL.

Rodrigues, J. (2017) "O que é o Processamento de Linguagem Natural?", https://medium.com/botsbrasil/o-que-%C3%A9-o-processamento-de-linguagem-natural-49ece9371cff.