Natural language question generation from Connected Open Data: a study of possibilities

Emilio Luis Faria Rodrigues
IFRR Instituto Federal de Educação, Ciência e Tecnologia de Roraima
Boa Vista, RR, Brasil
Emilio.luiz@ifrr.edu.br

Augusto Lopes da Silva, Sandro José Rigo, Denis Andrei Araújo,
UNISINOS
São Leopoldo, Brasil
augustols@outlook.com, rigo@unisinos.br,
denis.andrei.araujo@gmail.com

Abstract — Accelerated growth of Linked Open Data has been observed. There are several motives for this. Some situations are the generation of these bases from texts, another is the amount of open data generated from information systems. Their growth results in a large volume of information available. This fact brings the possibility of large-scale use in question-and-answer systems. Which have a general function that is based on a structure of information to generate phrases and to conference of produced answers. From literature studies, we foresee an opportunity of bigger use, in several applications. It is possible to generate natural language phrases from linked open databases, in Portuguese. In addition, we identified challenges to effective applicate linked open data resources for this purpose effectively. In this way, this work aims to see which aspects of the structure of open linked databases could be used as support in generating natural language questions.

Keywords— Linked Open Data; Natural Language Generation; Semantic Web.

I. INTRODUÇÃO

Alguns avanços tecnológicos recentes, tais como o aumento na capacidade de comunicação, processamento e armazenamento de dados definiram uma situação na qual a quantidade de dados gerados e compartilhados é muito grande e seu potencial de uso é muito promissor. Várias aplicações foram criadas a partir do uso destes recursos, tais como as aplicações de redes sociais, ferramentas de publicação de notícias, ambientes virtuais de aprendizagem, entre outras. A partir deste contexto é possível observar uma situação em que o número de dados disponíveis para os mais diversos setores da sociedade é muito grande, este volume cresce em ritmo acelerado. Esta situação vem estimulando iniciativas para melhor utilização destes dados, sendo que uma delas é a iniciativa conhecida como Dados Abertos Conectados [12].

A definição de Dados Abertos Conectados dada por [2] descreve em primeiro lugar o aspecto também conhecido como Dados Abertos, que está ligado com iniciativas governamentais para publicação e disponibilização de dados gerados no âmbito de atividades institucionais. Algumas iniciativas como a Lei de Acesso à informação (Lei 12.527, 18 novembro de 2011) promovem o suporte jurídico que estimula as instituições a também divulgarem constantemente os dados de suas atividades. Outro aspecto dos Dados Abertos está relacionado com iniciativas populares de divulgação de informação de forma espontânea e colaborativa, como no caso da Wikipedia [4].

A Web Semântica proporciona recursos para descrever dados de modo a possibilitar a automatização do seu uso e de sua integração. Recursos como ontologias e padrões de descrição de recursos vêm sendo amplamente utilizados com esta finalidade. Estas iniciativas são suportadas por setores em diversas áreas [1]. Um exemplo de melhorias proporcionadas com a Web Semântica é a iniciativa da DBPedia [3], projeto que objetiva automaticamente anotar as informações disponibilizadas na Wikipedia, em formado estruturado. Com base neste cenário, a iniciativa conhecida como Dados Abertos Conectados envolve não apenas o aspecto de acesso aos dados, que é proporcionado pelas iniciativas de Dados Abertos. Esta iniciativa destaca justamente o potencial para aplicações quando os dados também estiverem anotados segundo os padrões da Web Semântica. Além de estarem acessíveis, os dados estão representados de modo a serem usados por aplicações, diretamente [2].

O atual crescimento das bases de dados abertas e conectadas gerou um grande volume de informações. Alguns exemplos destas bases de dados abertas e conectadas são a DBpedia[1], YAGO3[2], Wikidata[3], entre outros. Informações preliminares sobre estas iniciativas estão descritas na Tabela 1.

TABELA 1 EXEMPLOS DE BASES DE DADOS ABERTAS E CONECTADAS

Nome	Descrição	Tamanho
DBpedia	Base de conhecimento com informações estruturadas da Wikipedia, tornando estas informações disponíveis na Web em formato anotado.	Mais de 16,9 milhões de entidades.
YAGO3	É uma enorme base de conhecimento semântica, derivada da Wikipedia, WordNet e GeoNames.	Mais de 10 milhões de entidades.
Wikidata	Projeto da Fundação Wikimedia: um banco de dados secundário, multilíngue, colaborativo e livre, que coleta dados estruturados para fornecer suporte à Wikipédia, ao Wikimedia Commons, aos demais projetos Wikimedia e muito mais.	Mais de 26 milhões de itens de dados.

[1] http://www.dbpedia.pt/
[2] http://yago-knowledge.org/
[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

Estas bases apresentam uma possibilidade de utilização em larga escala, como fonte de informações para sistemas de pergunta e resposta [13]. Os sistemas de pergunta e resposta possuem um funcionamento geral que está baseado em uma estrutura de informações a ser usada na geração de frases e na conferência das respostas. Alguns dos trabalhos nesta área são destinados a realizar uma transcrição entre as perguntas feitas em Linguagem natural pelos usuários e um formato de consulta, como SQL ou SPARQL [6]. Outros sistemas atuam também na geração das frases contendo as perguntas a serem realizadas [10]. Para estas atividades observa-se uma tendência no uso ontologias como fonte destas informações, tanto para a geração de perguntas como para consulta de respostas [8].

A partir do contexto descrito, observa-se uma situação em que existe um conjunto de recursos (as bases de dados abertos conectados) que está em franco crescimento e que pode ser utilizado para sistemas de perguntas e respostas, seja com finalidades gerais, seja para finalidades educativas. O desafio que se coloca para a sua utilização automática reside na definição de técnicas que permitam gerar automaticamente frases em linguagem natural, a partir dos dados abertos conectados, de modo a permitir que estas sejam usadas pelos sistemas de perguntas e respostas.

Os métodos para geração de linguagem natural necessitam alguns subsídios para sua execução [14], sendo que um destes subsídios consiste em informações para o planejamento do texto a ser gerado. Este problema é observado em diversos trabalhos estudados e em geral é resolvido com intervenção manual, o que é custoso e pode não ser eficiente em diversos contextos. Ocorre que nas bases de dados abertos conectados, parte dos recursos consiste exatamente neste tipo de subsídio, ou pode ser usado com esta finalidade. Ou seja, a estrutura interna de anotação dos dados nestas bases descreve exatamente um conjunto de relações que pode ser utilizada neste sentido. As bases de dados abertos conectados possuem uma estrutura e semântica definidas, com conceitos descritos e relacionados entre si. Desta forma, oferecem o suporte para os métodos de GLN, de modo que diversos trabalhos recentes [15; 16] indicam esta possibilidade como promissora e analisam alguns de seus desafios.

Este recurso pode ser utilizado como base para sistemas de pergunta e resposta, bem como para jogos educacionais. Alguns trabalhos já foram desenvolvidos com propósitos restritos a um determinado contexto, como é o caso do trabalho de [17], que apresenta uma aplicação multilíngue baseada em ontologias para informações de museus na Web. A abordagem baseia-se no uso da Web de dados ligados aplicados ao domínio do património cultural, tendo sido desenvolvido um aplicativo Web que usa ontologias da Web Semântica com técnicas de geração de linguagem natural para gerar descrições multilíngues coerentes sobre objetos de museus. Trabalho com objetivo similar pode ser observado em [18], com foco no uso de dados geográficos abertos para gerar automaticamente descrições para redes de sensores Hidrológicos.

Este artigo apresenta um estudo sobre as possibilidades de uso da estrutura de bases de dados abertos conectados para apoiar a descrição de um modelo que utilize esses dados disponíveis na web para montar frases contendo perguntas em linguagem natural. Durante as pesquisas realizadas, observou-se a oportunidade de maior utilização, em aplicações diversas, das possibilidades de geração de frases em linguagem natural a partir de uma base de dados abertos conectados, na língua portuguesa. Desta forma, esse artigo visa responder as seguintes questão de pesquisa: Quais os aspectos da estrutura de bases de dados abertas conectadas que podem ser utilizados como apoio na geração de perguntas em linguagem natural?

O objetivo geral do trabalho é definir um modelo que utiliza bases de dados abertos e conectados para gerar frases com perguntas em linguagem natural, avaliando as características e potencialidades dos recursos de web semântica, em especial no caso de dados abertos e conectados, para uso na geração de perguntas em linguagem natural, modelando um mecanismo de geração automática de perguntas usando dados abertos e conectados e descrevendo uma aplicação para validar experimentalmente os mecanismos desenvolvidos.

O restante do texto apresenta os trabalhos relacionados estudados, a metodologia adotada e arquitetura proposta, avaliações realizadas e conclusões.

II. TRABALHOS RELACIONADOS

O trabalho de Menezes [5] apresenta um sistema de perguntas e respostas que utiliza ontologias e técnicas de recuperação de informações na análise da pergunta, além de usar palavras chaves com o uso da combinação de ontologias com agentes de softwares. O AskNow [6] é um framework através do qual o usuário pode colocar consultas em Inglês em uma base de conhecimento RDF e traduzir estas para consultas SPARQL.

SPARQL2NL [7] é um framework que permite verbalizar consultas utilizado o SPARQL em uma base de conhecimento como a DBpedia e convertendo-as em linguagem natural. O framework normaliza a consulta e extrai informações, processando a representação genérica gerada através de uma consulta, aplicando regras de redução e substituição para melhorar a legitimidade da verbalização e transformando toda a informação gerada em representação final da linguagem natural.

PortNLG [9] é um sistema de realização superficial para geração de textos em português, baseado em regras para o português brasileiro, que trata da tarefa de linearização sentencial para aplicações computacionais que necessitem apresentar dados de saída em formato textual. O RealText-lex [19] é um framework que tem como objetivo construir uma estrutura de padrões para lexicalização de triplas de RDF extraídos da DBpedia. O framework tem quatro módulos orientados para a geração de padrões de lexicalização.

NaturalOWL [8] é um sistema de geração de linguagem natural que produz textos que descrevem os indivíduos ou classes de ontologias OWL. No sistema, podem-se publicar informações em OWL na Web, juntamente com textos produzidos automaticamente correspondentes em vários idiomas. No trabalho de Ell e Harth [10] o sistema atua na utilização de recursos de dados abertos e conectados para geração de frases em linguagem natural utilizando templates com aspectos que relacionam os elementos de dados conectados além de elementos de lexicalização.

No trabalho de Dumas e Klein [11], o sistema propõe uma arquitetura de geração de linguagem natural a partir de dados vinculados que automaticamente aprende com modelos de sentenças e planejamento de documentos estatísticos a partir de conjuntos RDF e textos paralelos. O objetivo do sistema é gerar descrições curtas, equivalentes a entradas da Wikipédia, de entidades encontradas em conjuntos de dados vinculados.

Por meio do estudo dos trabalhos relacionados apresentados anteriormente, realizou-se um processo de identificação dos principais desafios da área e também análise da experimentação de soluções constatadas. Foram identificadas, nestes trabalhos, as principais técnicas atuais para geração de frases, sendo que neste aspecto pode-se observar um conjunto de abordagens que vai desde o uso de padrões textuais até recursos linguísticos e aprendizagem de máquina. Ainda não são observados trabalhos consolidados para geração de perguntas de forma automática em português, bem como não se observou a existência de um estudo aprofundado sobre os dados abertos conectados para identificar relacionamentos que possam ser de utilidade para a geração de frases. Sendo assim, de acordo com os objetivos deste trabalho, estas lacunas serão abordadas a partir do desenvolvimento de um estudo sobre padrões de dados abertos e conectados e características que podem ser de interesse para a geração de frases contendo perguntas em linguagem natural.

TABELA 2 – COMPARAÇÃO ENTRE TRABALHOS ESTUDADOS

Artigos	**SPARQL**	**Lexicali-zação**	**Dados Abertos**	**Dados Conectados**
(NOGMO et al., 2013)	SIM	NÃO	SIM	SIM
(DOUGLAS et al., 2013)	NÃO	SIM	NÃO	NÃO
(ADROUTSOPOULOS et al., 2013)	NÃO	SIM	NÃO	NÃO
DUMAS e KLEIN, 2013)	NÃO	NÃO	SIM	SIM
(PERERA e NAND, 2016)	NÃO	SIM	SIM	SIM
ELL e HARTH, 2016)	SIM	NÃO	SIM	SIM

Na tabela 2 estão agrupados os atributos analisados nos sistemas de geração de linguagem que foram alvo deste estudo. Por meio do estudo dos trabalhos relacionados apresentados anteriormente, realizou-se um processo de identificação dos principais desafios da área e também análise da experimentação de soluções constatadas. Foram identificadas, nestes trabalhos, as principais técnicas atuais para geração de frases, sendo que neste aspecto pode-se observar um conjunto de abordagens que vai desde o uso de padrões textuais até recursos linguísticos e aprendizagem de máquina.

Ainda não são observados trabalhos consolidados que realizaram um estudo aprofundado sobre os dados abertos conectados para identificar relacionamentos que possam ser de utilidade para a geração de frases. Sendo assim, de acordo com os objetivos deste trabalho, estas lacunas serão abordadas a partir do desenvolvimento de um estudo sobre padrões de dados abertos e conectados e características que podem ser de interesse para a geração de frases contendo perguntas em linguagem natural.

III. METODOLOGIA E MODELO PROPOSTO

Neste trabalho, com vistas à alcançar os objetivos propostos, uma das etapas desenvolvidas estava destinada à análise detalhada dos dados abertos conectados no que se refere aos elementos que podem ser usados na geração de linguagem natural. Outra etapa está relacionada com a descrição de um modelo para apoiar a prototipação de uma ferramenta que permita realizar a geração automática de frases a partir das bases de dados conectadas.

O potencial para geração de perguntas que uma das maiores bases de dados abertos e conectados, a DBpedia, possui pode ser avaliado a partir do estudo de seu contexto. O projeto da DBPedia possui como objetivo a transcrição dos dados em formato não estruturado (formato textual) encontrados em ambientes de divulgação de dados em geral na Web, sendo que neste caso foi usado, em particular, o exemplo da Wikipedia. A Figura 1 abaixo ilustra o processo envolvido no uso da Wikipedia para a geração de dados para a DBPEDIA. Parte-se de um texto em língua natural na Wikipedia e o resultado final é um documento em formato RDF na DBPEDIA. A figura 1 ilustra partes de uma página web da Wikipedia sobre a cidade de Berlin, que pode ser acessada no endereço https://en.wikipedia.org/wiki/Berlin.

A parte inferior da Figura 1 ilustra diversos tipos de dados que podem ser extraídos dos conjuntos de dados em texto não estruturado. Um exemplo destes dados é a informação de latitude e longitude, que são compostas por valores numéricos (wgs84:lat 52.500577; wgs84:long 13.398889). Outro exemplo é a informação de rótulos (labels) usados para descrever textualmente aspectos do conceito representado, neste caso em duas línguas, inglês e italiano (rdfs:label “Berlin”@en; “Berlino”@it;). Um terceiro exemplo são os conceitos que relacionam Berlin (conceito geral da página) com os conceitos de cidade e outros conceitos complementares a este, como o conceito de local muito populoso (dbpedia:Berlin rdf:type dbpedia-owl:City , sbpedia-owl:PopulatedPlace).

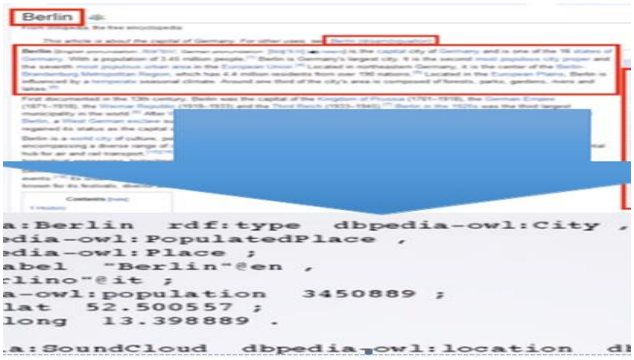

Figura 1 – Equivalência de textos em língua natural (Wikipedia) e dados no formato RDF (DBpedia)

A partir destes dados anotados nas bases de dados abertos conectados como o exemplo da DBpedia, são vislumbradas possibilidades de geração automática ou semiautomática de frases curtas com perguntas em linguagem natural. As relações e sua semântica serão empregadas como elemento para auxiliar e guiar o planejamento da geração das frases. A partir de locais de consulta com formatos SPARQL, tais como o seguinte endereço (https://dbpedia.org/sparql), é possível o acesso e realização de consultas nesta linguagem.

A seguir estão relacionados alguns exemplos de resultados obtidos nestas consultas SPARQL, tendo como foco o conceito "Porto Alegre" e duas relações associadas com este conceito na ontologia utilizada na DBPedia (populationTotal e country). Para o conceito ""Dbpedia: Porto Alegre"" podem ser obtidos exemplos de relações e valores como: ""Dbpedia-owl: populationTotal 15099399"" ou ""Dbpedia-owl: country dbpedia: Brazil"".

Em uma interpretação livre das relações indicadas, seria possível considerar a geração de exemplos de perguntas em linguagem natural que podem ser geradas, tais como "Qual é a população total de Porto Alegre?", para a primeira relação, ou então "Em que país fica Porto Alegre?" para a segunda relação.

A. Desafios observados

A primeira fonte de informação escolhida para este trabalho foi a DBpedia. A metodologia a seguir foi definida para esta primeira etapa: estudar as descrições da ontologia da DBpedia para identificar quais os itens possuem elementos com viabilidade para auxiliar na automatização da geração de frases com perguntas a partir de dados abertos e conectados.

A análise inicial partiu do pressuposto de que as relações definidas para os elementos da ontologia estudada possuem aspectos que descrevem uma semântica bem definida e que podem ser utilizados para o processo de geração de linguagem. Além da descrição das relações, estima-se ser possível utilizar sua qualificação, como tipo de dados, intervalos de valores, ou outras restrições.

Figura 2 - Conceito sobre população total na DBpedia. Fonte: Dbpedia

A partir da Figura 2, alguns aspectos preliminares são analisados e comentados. Estes foram descritos na Tabela 3 abaixo.

TABELA 3 – ALGUNS ASPECTOS ESSENCIAIS IDENTIFICADOS

Elemento	Significado
rdfs: label	Contém o rótulo da relação. Pode ser usado para gerar a frase. Sinônimos do elemento "label" também podem ser usados.
rdfs: range	Indica o tipo de dados que será usado para completar a relação. Pode ser usado para apoiar a correção da resposta dada à pergunta.
owl: sameAs	Relação que indica uma existência de outra relação que tem o mesmo significado. Permite considerar o uso dos termos da relação indicada como apoio na geração de frases.
rdfs: type: datatype property	Relação que descreve literais como valores e, portanto, pode também indicar tipos específicos de dados e variações destes valores. Todos estes aspectos podem ser utilizados na geração de frases.
rdfs: type: object property...	Relação que descreve uma conexão com outros conceitos. Permite identificação de conjuntos restritos de valores.

Com base na metodologia definida, foi ampliada a quantidade de elementos estudados, sendo que uma parte destes está resumida na Tabela 4. O total de elementos analisados até o momento possibilita identificar algumas tendências e sinalizar possibilidades de aplicação de algoritmos para aproveitar estas situações na geração de frases curtas para perguntas.

TABELA 4- TRECHO PARCIAL DA TABELA DE ANÁLISE, PARA O CONCEITO HTTP://DBPEDIA.ORG/PAGE/PORTO_ALEGRE

Propriedade / Recurso	**Property**	**Value**
dbo:PopulatedPlace/area	rdf:type	owl:DatatypeProperty
	rdfs:domain	dbo:PopulatedPlace
	rdfs:label	área (km^2)
	rdfs:range	http://dbpedia.org/datatype/squareKilometre
	Graph	http://dbpedia.org/resource/classes#
dbo:PopulatedPlace/areaTotal	rdf:type	owl:DatatypeProperty
	rdfs:domain	dbo:PopulatedPlace
	rdfs:label	área toral (km^2)

	rdfs:range	http://dbpedia.org/datatype/squareKilometre
	Graph	http://dbpedia.org/resource/classes#
dbo:PopulatedPlace/populationDensity	rdf:type	owl:DatatypeProperty
	rdfs:domain	dbo:PopulatedPlace
	rdfs:label	population density (/sqkm) (en)
	rdfs:range	http://dbpedia.org/datatype/inhabitantsPerSquareKilometre
	Graph	http://dbpedia.org/resource/classes#
dbo:ontology/country	rdf:type	rdf:Property; owl:ObjectProperty
	rdfs:comment	The country where the thing is located. (en)
	rdfs:label	country (en)
	rdfs:range	dbo:Country
dbo:ontology/isPartOf	rdf:type	rdf:Property; owl:ObjectProperty
	rdfs:label	is part of (en)
dbo:ontology/populationTotal	rdf:type	owl:FunctionalProperty; rdf:Property; owl:DatatypeProperty
	rdfs:label	population total (en); população total (pt)
	rdfs:range	xsd:nonNegativeInteger

A análise prevista nesta etapa parte, portanto, do uso de recursos de lexicalização já definidos nas bases de dados abertos e conectados. O primeiro elemento investigado é o elemento "label". Para este elemento foram realizados testes de consultas para verificar o seu conteúdo, além de realizar a sua validação, quanto à possibilidade de utilização na lexicalização das frases, com apoio de profissionais especializados. Como segundo elemento investigado, foram catalogados e analisados os exemplos de relações que permitem a geração automática de lexicalização, seja de forma automática, com recursos de processamento de linguagem natural, seja com anotação manual. Nestes casos, a validação destes elementos como possibilidade permite a inclusão e constante atualização de uma etapa ou opção no algoritmo de geração de perguntas em linguagem natural.

Um terceiro elemento considerado são as definições conjuntas de tipo (rdf:type) e de faixa de valores (rdf:range), que foram considerados, nesta análise, como potenciais elementos para apoiar a definição de alguns dos aspectos de perguntas, tais como o tipo de pergunta e os valores possíveis. Por exemplo, uma relação cuja faixa de valores está definida como valores inteiros permite considerar este aspecto na geração da pergunta e também em sua posterior avaliação.

Neste sentido, pode-se considerar um exemplo ilustrativo. A partir de um conceito "dbo:ontology/populationTotal" indicado na tabela 2, pode ser identificado o uso do valor disponível na relação "label", bem como sinônimos destes valores, para a geração de frases significativas para geração de perguntas sobre este conceito. Portanto o ponto inicial para a geração de frases é o conceito da DBPedia: "Dbpedia-owl: populationTotal".

A estimativa de geração de perguntas relacionadas pode ser, por exemplo: "Qual é a população total de ..."; "Me diga qual é a população total da cidade de ..."; "Diga qual é a quantidade de pessoas que vivem em ...". Os elementos disponíveis diretamente para uso são a semântica do elemento rdfs:label, que contém uma descrição textual do conceito associado, pois o termo "população total" está representado como valor do elemento rdfs:label. Outros elementos disponíveis indiretamente são os sinônimos dos termos do elemento rdfs:label. Por exemplo, o termo "população total" pode ser considerado um termo sinônimo de "quantidade de pessoas que vivem em".

B. *Algoritmo e arquitetura*

Seguindo a linha geral de trabalhos como [10, 11], diversos recursos foram descritos para apoiar a geração das frases. Alguns dos recursos são padrões definidos manualmente, com base no estudo das relações das bases de dados analisadas, contendo aspectos de lexicalização e de substituição de valores. Outros padrões são mais elaborados e utilizam as relações semânticas detectadas, para então definir o tipo de pergunta a ser gerada e os elementos de lexicalização a serem usados. Além disso, estão previstos recursos adicionais para compor um cenário de flexibilidade de geração de frases, com uso de variações linguísticas, uso de sinônimos e termos compostos.

Os elementos inicialmente definidos iniciam pelos padrões gerais de pergunta com aspectos de lexicalização, que associam de modo manual uma relação e conceito da base de dados aberta e conectada com um termo a ser usado para a geração das frases. Estes padrões possuem o formato descrito na figura 3.

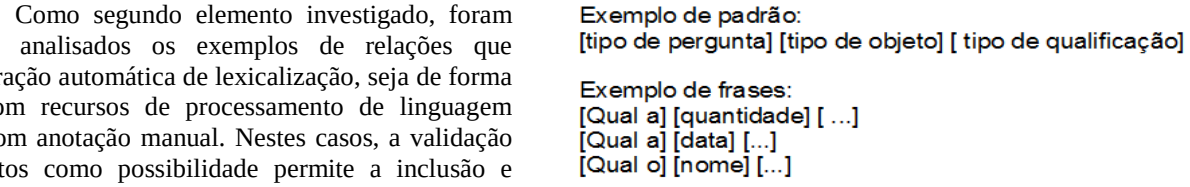

Figura 3- Exemplos de padrões de lexicalização

O segundo grupo de elementos é descrito a partir de uma associação de relações com tipos de pergunta, sendo que as perguntas, nestes casos, são geradas de forma mais flexível. Nestes casos são mapeadas preliminarmente associações entre relações e as perguntas respectivas, levando em conta a semântica da relação e atributos, como os seus tipos de dados, por exemplo. Nestes casos, alguns exemplos são, para a relação do tipo ""total"", as perguntas do tipo ""qual", ou "quanto"". Ou então, para as relações do tipo "" Data"", as perguntas do tipo ""quando"".

A diferença fundamental para o primeiro grupo de padrões é que este padrão não é gerado manualmente, mas gerado partir de um conjunto de relacionamentos entre padrões de lexicalização para as perguntas e um conjunto de relações associado.

O terceiro grupo de elementos está definido partir da experimentação de uso dos elementos analisados nas bases de dados abertos e conectados. Por exemplo, um estudo inicial realizado permitiu identificar que o elemento "label" possui padrões de lexicalização com potencial de uso. Deste modo, este tipo de estudo será continuado e serão catalogados elementos e sua possibilidade de utilização. De modo a fomentar a automatização sempre que possível estes padrões são definidos com base na listagem das relações, dos itens a serem utilizados e também em operações a serem realizadas, apoiando assim o ajuste dos termos sempre que necessário.

A dinâmica geral do algoritmo adotado, portanto, considera as seguintes etapas: a) Seleção do conjunto de bases de dados abertos a utilizar; b) Seleção do tipo de conceito desejado pelo usuário; c) Realização da consulta SPARQL para recuperar as relações e valores associados com o conceito pesquisado; d) Análise do conjunto de relações obtido com a consulta; e) Utilização do conjunto de padrões de lexicalização para a geração das frases curtas.

Para apoiar a execução desta abordagem proposta, os componentes necessários estão relacionados na arquitetura geral, resumida na Figura 4. Existem duas formas de interação com o sistema, sendo uma de parte dos usuários, que irão utilizar o sistema para consultar conceitos e receber as frases com as perguntas em linguagem natural. A segunda forma de interação é por parte de especialistas, responsáveis pela identificação e documentação dos padrões de lexicalização a serem empregados.

Durante a operação prevista para o atendimento do usuário desta arquitetura, os módulos de número 1 e 9 consistem apenas de elementos dedicados à interface do sistema com os usuários para permitir a escolha de conceitos pelo mesmo e a exibição das frases geradas como resultado do processo e 3, 5 e 7 dedicados ao atendimento do usuário Especialista ou Administrador. O módulo identificado como item 2 possui como objetivo a tradução da seleção de um conceito em uma consulta SPARQL para ser realizada em uma base de dados aberta e conectada.

Os componentes de Mapeamento de Padrões e de Análise de relações utilizam os aspectos do algoritmo definido para permitir que sejam relacionadas diferentes relações com o seu correspondente item lexical, compondo assim um banco de informações necessárias para a geração das frases. O Item de Mapeamento de padrões (item 5) é o responsável por armazenar um conjunto de informações sobre as relações analisadas na base de dados abertos e conectados e que foram consideradas como viáveis para a utilização na geração de linguagem natural. O item de Análise de relações (item 4) é o item responsável por realizar a verificação das relações obtidas na consulta realizada pelo usuário sobre um determinado conceito e identificar se o conjunto de padrões já selecionados possui alguma destas relações.

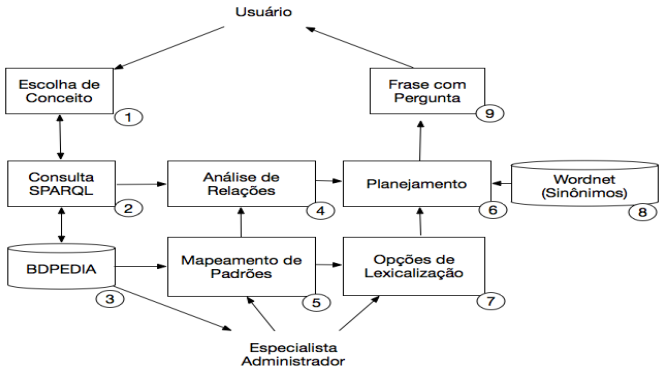

Figura 4- Visão geral da Arquitetura

Por fim os elementos de Opções de Lexicalização (item 7) e de Planejamento (item 6) são os responsáveis pela integração final das informações. O elemento de Opções de Lexicalização mantém uma base de padrões para serem usados na geração das frases. Já o item de Planejamento possui como tarefa a integração dos aspectos lexicais com os dados obtidos com a consulta SPARQL feita sobre o conceito escolhido, junto com recursos adicionais, tais como o Wordnet, que permitem a expansão das etapas de lexicalização, com uso de recursos linguísticos.

IV. AVALIAÇÃO E CONTRIBUIÇÕES

Os aspectos considerados como centrais neste trabalho, em termos de contribuição, estão associados com o estudo preliminar do atual contexto ligado às bases de dados abertas e conectadas, bem como o estudo das relações descritas nestas bases e na sua semântica. Desta forma, buscou-se avançar principalmente nestes aspectos do estudo.

Como forma de avaliação sobre os possíveis resultados destes subsídios identificados, foi realizado no escopo do trabalho uma implementação da arquitetura proposta, que proporcionou os elementos necessários para uma etapa de geração de frases curtas com perguntas em linguagem natural e também permitiu a avaliação destes resultados.

Foram desenvolvidas experimentações com o uso de NodeJs e com o uso de Java, para realizar as etapas do processo geral descrito. Estes experimentos permitiram validar os procedimentos de acesso às bases de dados abertas e conectadas, tendo sido usada como caso de estudo a DBPedia, tanto para busca de dados disponíveis em língua portuguesa como para dados em língua inglesa. O principal objetivo foi a implementação do acesso e recuperação de dados, seguido pelo posterior tratamento destes para a geração de frases com perguntas em linguagem natural e, por fim, a sua avaliação crítica por parte de uma especialista em linguística.

Foi utilizado no experimento uma parte do conjunto de relações estudadas. Definiu-se, por conveniência de acesso e facilidade de obtenção de maior quantidade de dados, a busca de conceitos em língua inglesa. Os mesmos procedimentos podem ser utilizados em outras línguas, dentro de um escopo de restrições de alguns de seus aspectos. Deste modo, definiu-se o uso de perguntas do tipo "Who", "How

much" e "What". Estes tipos iniciais de perguntas definidos estão associados com aspectos da semântica de relações estudadas. Desta forma, o primeiro caso está associado com a identificação de uma pessoa, o segundo está associado com relações descrevendo quantidades e o último está associado com outras situações gerais, como forma de avaliar resultados não mapeados previamente.

No caso da execução de uma consulta com uso da linguagem SPARQL, o resultado típico permite que sejam consultadas diversas relações para o conceito que foi o motivo da consulta. A figura 5 ilustra, por exemplo, o resultado de uma consulta para o termo cidade (city) e as diversas relações associadas com um conceito como "cidade", dentre as quais é possível identificar as relações como "label', ou "range', entre diversas outras. Quando a consulta é realizada para a busca do nome de uma pessoa, por exemplo, o mesmo mecanismo de identificação de relações possibilita a descrição de relações que identificam o tipo e domínio, indicando o pertencimento à classe pessoa. Da mesma forma, em diversas outras categorias de conceitos, existem relações apropriadas para a sua identificação, que podem ser então utilizadas na geração de subsídios para a geração de perguntas em linguagem natural.


```
{
"http://dbpedia.org/ontology/city" :
{ "http://www.w3.org/1999/02/22-rdf-syntax-ns#type" :
[ { "type" : "uri", "value" : "http://www.w3.org/2002/07/owl#ObjectProperty" } ,
{ "type" : "uri", "value" : "http://www.w3.org/1999/02/22-rdf-syntax-ns#Property" }
] ,
"http://www.w3.org/2000/01/rdf-schema#subPropertyOf" :
[ { "type" : "uri", "value" :
"http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#hasLocation" }
] ,
"http://www.w3.org/2002/07/owl#equivalentProperty" :
[ { "type" : "uri", "value" : "http://www.wikidata.org/entity/P131" }
] ,
"http://www.w3.org/2000/01/rdf-schema#label" :
[ { "type" : "literal", "value" : "\u03C0\u03CC\u03BB\u03B7" , "lang" : "el" } ,
{ "type" : "literal", "value" : "ville" , "lang" : "fr" } ,
{ "type" : "literal", "value" : "cathair" , "lang" : "ga" } ,
{ "type" : "literal", "value" : "Stadt" , "lang" : "de" } ,
{ "type" : "literal", "value" : "city" , "lang" : "en" } ,
{ "type" : "literal", "value" : "miasto" , "lang" : "pl" } ,
{ "type" : "literal", "value" : "stad" , "lang" : "nl" }
] ,
"http://www.w3.org/2000/01/rdf-schema#range" :
[ { "type" : "uri", "value" : "http://dbpedia.org/ontology/City" }
] ,
"http://www.w3.org/2002/07/owl#sameAs" :
[ { "type" : "uri", "value" : "http://dbpedia.org/ontology/city" }
] ,
"http://www.w3.org/ns/prov#wasDerivedFrom" :
[ { "type" : "uri", "value" :
"http://mappings.dbpedia.org/index.php/OntologyProperty:city" } ] }
}
```


Figura 5 - Trecho em Jason para descrição parcial de resultado de consulta - Fonte: dbpedia.org, w3.org

Na geração deste teste prevendo a avaliação, foram usadas algumas destas relações especificamente. Deste modo, quando é realizada uma consulta e o elemento que representa o sujeito da tupla recebida na consulta pode ser verificado como possuindo a relação do tipo "dbo:domain", então esta informação pode ser utilizada com confiança como o delimitador para o tipo de questão a ser associada. Por exemplo, se a relação de domínio (dbo:domain) apontar para uma pessoa como domínio, a questão pode ser realizada com segurança a respeito deste aspecto. O mesmo ocorre com as demais relações mapeadas.

Para a avaliação do protótipo procedeu-se à realização de experimentos com a finalidade de identificar os aspectos positivos gerados pelo estudo das relações estudadas. Um dos aspectos de interesse foi a funcionalidade do protótipo na busca de tratamento dos dados. O outro aspecto foi a sua capacidade de geração de frases com perguntas em linguagem natural. Para a avaliação dos resultados obtidos foi executada uma análise por especialistas em linguística, buscando identificar problemas de lexicalização e de uso da linguagem.

Como forma de diversificar as possibilidades de obter subsídios para a análise do procedimento desenvolvido, foram definidos elementos de consulta que representam categorias diversificadas. Por exemplo, foram identificadas categorias como pessoas ("Bill Gates", "Barack Obama"), cidades ("Chicago", "Porto alegre") e outras entidades como animais ("lion", "tiger"), doenças ("Influenza") e eventos ("World War II", "French Revolution").

Com base neste conjunto de tópicos escolhidos e nos três tipos de perguntas elencados inicialmente, foram geradas as frases curtas com perguntas e depois as mesmas foram analisadas por um especialista em linguística. Um ponto de vista inicial da análise foi identificar se a frase está correta gramaticalmente. Outro ponto de vista de análise foi buscar identificar a correção no uso da linguagem da forma como foi gerada a pergunta. Estes dois aspectos estão relacionados intimamente com o mapeamento das relações disponíveis nas bases de dados abertas e conectadas e a sua semântica. Melhorias nestes estudos levam a melhorias nos resultados obtidos. Por exemplo, uma mesma pergunta pode ser formulada corretamente com diferentes estilos de escrita, o que precisa ser identificado como possibilidade a partir do estudo das relações e de informação gerada pelos especialistas em linguística.

Foram geradas 205 questões, com base nos tópicos definidos. A análise dos resultados por especialista em linguística permitiu identificar situações que serão tratadas em trabalhos futuros. Além da identificação de correção das sentenças, esta análise foi realizada para indicar também outras formas de gerar as frases com as perguntas, o que poderá servir de subsídio futuro para a diversificação dos resultados.

Uma das situações a destacar na avaliação é a existência de relações contendo descrições que não são adequadas para a geração de perguntas, pois indicam elementos adicionais, por vezes técnicos, dos conceitos tratados. Por exemplo, uma destas é a situação em que as relações são usadas para identificar termos adicionais sobre material a respeito do conceito. Por exemplo, é comum encontrar-se o termo "has abstract" associado com diversos conceitos indicando um texto adicional sobre aquele conceito. No caso de um conceito associado com uma pessoa, não terá sentido a realização de uma pergunta sobre esta relação. Ou então, para alguns conceitos, especialmente animais ou fenômenos atmosféricos, é comum a existência de relações contendo imagens sobre os mesmos, nas quais os termos usados são tais como "image', "image width", "image caption'. Ou então, em mais um exemplo de interesse, conceitos descrevendo doenças podem conter relações apontando para recursos como bases de dados nas quais existem mais informações sobre as mesmas, como "DiseasesDb".

Além desta identificação, os resultados possibilitaram identificar situações de correção gramatical das frases. Na figura 6 estão resumidas algumas das análises realizadas e seus resultados. Todas as situações de interesse identificadas podem ser observadas nesta figura. A figura 6 representa alguns dos termos de consulta acessados na Base de dados

aberta, na sua primeira coluna. Neste caso, estão representados os termos “French Revolution”, “World War II”, “Porto Alegre”, “Barack Obama”. Para cada um destes termos, a primeira coluna mostra as frases geradas. As frases corretas estão representadas em letras pretas e as frases incorretas estão representadas em letras vermelhas.

A segunda coluna, na figura 6, exibe um extrato de alguns exemplos de resultados das correções e avaliações realizadas, sendo que podem ser identificadas tanto as correções das frases, quando necessário, como exemplos das mesmas frases contendo formulação diferente, como exemplo de diversidade na geração dos resultados.

French Revolution	CORRETA
What's French Revolution's has abstract?	What's the French Revolution history?
What's French Revolution's date?	
What's French Revolution's Event Name?	
What's French Revolution's image caption?	
World War II	CORRETA
What's World War II's has abstract?	What's the World War II history?
What's World War II's causalties?	What were the WWII effects?
What's World War II's combatant?	Who were the WWII soldiers?
What's World War II's place of military conflict?	Where did the WWI take place?
What's World War II's result?	
Porto Alegre	CORRETA
What's Porto Alegre's area (km2)?	What's Porto Alegre's area?
What's Porto Alegre's area total (km2)?	
What's Porto Alegre's population density (/sqkm)?	
What's Porto Alegre's has abstract?	
How much is Porto Alegre's area (m2)?	
What's Porto Alegre's area code?	
How much is Porto Alegre's area total (m2)?	
Barack Obama	CORRETA
What's Barack Obama's birth place?	What's Barack Obama's birthplace?
What's Barack Obama's birth date?	What's Barack Obama's birth date?
What's Barack Obama's name?	What's Barack Obama's name?
What's Barack Obama's has abstract?	What's Barack Obama's history?

Figura 6 - Resumo de resultados

Além destes aspectos citados, que priorizam a avaliação da correção das frases, do ponto de vista gramatical ou do ponto de vista de utilização da linguagem, deve ser destacado que o resultado desta avaliação foi considerado positivo do ponto de vista da funcionalidade. A abordagem avaliada permite que sejam acessados e tratados os dados oriundos de bases de dados abertas e conectadas, de forma eficiente, levando em conta tanto o acesso remoto, como a possibilidade de acesso local, tendo em vista que diversas destas permitem a sua cópia para utilização em contextos com necessidades maiores de performance.

V. Conclusões

O trabalho apresentado descreve a pesquisa realizada para explorar a possibilidade de geração de linguagem natural a partir de bases de dados abertas e conectadas. Foram identificadas potencialidades nesta área, de modo a ampliar o aproveitamento dos atuais conjuntos de dados disponíveis. Estes conjuntos possuem uma tendência de crescimento continuado. Além disso apresentam aspectos que favorecem o seu uso nesta tarefa de geração de linguagem natural, tais como a sua estruturação formal, a partir de ontologias ou vocabulários bem definidos. Estes conjuntos também apresentam como aspecto relevante a sua tendência de atualização constante, em especial aqueles conjuntos associados com iniciativas de governo aberto ou então conjuntos associados com redes sociais e de colaboração, tais como a Wikipedia, por exemplo.

Foram analisadas bases de dados abertas e conectadas no que diz respeito às relações descritas no formalismo RDF, o que possibilitou a indicação de procedimentos para o uso destas relações na geração automática de frases em linguagem natural, contendo perguntas que utilizam estas relações. Também foi descrito um modelo geral para a implementação de procedimentos de utilização destas relações para a geração de perguntas. Aspectos deste modelo foram validados e avaliados, com um protótipo que permitiu identificar aspectos positivos e melhorias a serem realizadas em atividades futuras.

A contribuição apresentada neste trabalho está associada com o estudo amplo dos atuais conjuntos de dados abertos e conectados e de suas relações. Espera-se, a partir desta abordagem, terem sido definidos passos na direção de um procedimento que auxilie a identificar elementos para apoiar as tarefas de geração de linguagem natural. A partir deste estudo, espera-se, portanto, qualificar os métodos de geração de frases em linguagem natural, de forma a ampliar o coeficiente de possibilidades de automatização desta atividade, pois atualmente os trabalhos estudados possuem um grau elevado de utilização de padrões lexicais definidos de modo manual.

Agradecimentos

O presente trabalho foi desenvolvido com o apoio da CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil, e as seguintes instituições: UNISINOS e IFRR.

VI. Referências

[1] HASLHOFER, Bernhard; ISAAC, Antoine. data. europeana. eu: The europeana linked open data pilot. In: International Conference on Dublin Core and Metadata Applications. 2011. p. 94-104.

[2] ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados Abertos Conectados: Em busca da Web do Conhecimento. Novatec Editora, 2015.

[3] LEHMANN, Jens et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, v. 6, n. 2, p. 167-195, 2015.

[4] Paletta, Francisco Carlos, and Marcos Luiz Mucheroni. "O desenvolvimento da WEB 3.0: Linked Data e DBPEDIA." Prisma. (2017).

[5] MENEZES, Crediné Silva de. AMORIM, Marta Talitha Carvalho Freire; CURY, Davidson; Sobre a aplicação de ontologias para orientar agentes a responder perguntas. Revista brasileira de informática na educação. Florianópolis. Vol. 22, n. 3 (2014), p. 1-12, 2014.

[6] DUBEY, Mohnish et al. AskNow: A Framework for Natural Language Query Formalization in SPARQL. In: International Semantic Web Conference. Springer International Publishing, 2016. p. 300-316.

[7] NGONGA NGOMO, Axel-Cyrille et al. SPARQL2NL: verbalizing sparql queries.

In: Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013. p. 329-332.

[8] ANDROUTSOPOULOS, Ion; LAMPOURAS, Gerasimos; GALANIS, Dimitrios. Generating natural language descriptions from OWL ontologies: the NaturalOWL system. Journal of Artificial Intelligence Research, v. 48, p. 671-715, 2013.

[9] Douglas Fernandes Pereira; DE NOVAIS, Eder Miranda; PARABONI, Ivandré. DA SILVA JUNIOR, Um Sistema de Realização Superficial para Geração de Textos em Português. Revista de Informática Teórica e Aplicada, v. 20, n. 3, p. 31-48, 2013.

[10] ELL, Basil; HARTH, Andreas. A language-independent method for the extraction of RDF verbalization templates. INLG-2014, 2014.

[11] DUMA, Daniel; KLEIN, Ewan. Generating natural language from linked data: Unsupervised template extraction. Association for Computational Linguistics, Potsdam, Germany, p. 83-94, 2013.

[12] HEATH, Tom; BIZER, Christian. Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, v. 1, n. 1, p. 1-136, 2011.

[13] SHEKARPOUR, Saeedeh et al. Question answering on linked data: Challenges and future directions. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016. p. 693-698.

[14] Reiter, E. B., & DALE, R. (1997). Building Applied Natural-Language Generation Systems. *Natural Language Engineering*, *3*, 57.

[15] Staykova, Kamenka. (2014). Natural Language Generation and Semantic Technologies. Cybernetics and Information Technologies. 14. 10.2478/cait-2014-0015.

[16] Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J. & Ngonga Ngomo, A.-C. (2016), 'Survey on Challenges of Question Answering in the Semantic Web', Semantic Web Journal .

[17] DANNÉLLS, Dana et al. Multilingual online generation from semantic web ontologies. In: Proceedings of the 21st International Conference on World Wide Web. ACM, 2012. p. 239-242.

[18] MOLINA, Martin; SANCHEZ-SORIANO, Javier; CORCHO, Oscar. Using open geographic data to generate natural language descriptions for hydrological sensor networks. Sensors, v. 15, n. 7, p. 16009-16026, 2015.

[19] PERERA, Rivindu; NAND, Parma; KLETTE, Gisela. RealText-lex: A Lexicalization Framework for RDF Triples. The Prague Bulletin of Mathematical Linguistics, v. 106, n. 1, p. 45-68, 2016.