Consulta multilíngue em banco de dados relacionais: Uma revisão sistemática

Robson C. Vieira[1], **João C. da Silva**[1]

[1]Instituto de Informática - Universidade Federal de Goiás (UFG)
Caixa Postal 131 – CEP 74.001-970 – Goiânia – GO – Brazil

{robsoncardoso,jcs}@inf.ufg.br

Abstract. *To retrieve multilingual information stored in relational databases, the initial query has traditionally been rewritten in other languages in order to achieve the desired results. This article presents a systematic review of the preprocessing methods and translation of the query. The multilingual query enables a person with no knowledge of the structured query language (SQL) or database structure to query queries using natural language or keywords in an L1 language and obtain results in the L2 language. In this way, one can, for example, conduct consultations in Portuguese and obtain results in another language. We considered the publications of the last five years (2015-2019) indexed by six internationally recognized scientific bases.*

Resumo. *Para recuperar informações multilíngues armazenadas em bancos de dados relacionais, tradicionalmente a consulta inicial precisa ser reescrita em outros idiomas a fim de obter os resultados pretendidos. Este artigo apresenta uma revisão sistemática dos métodos de pré-processamento e tradução da consulta. A consulta multilíngue possibilita que uma pessoa sem conhecimento da linguagem de consulta estruturada (SQL) ou da estrutura do banco de dados submeta consultas utilizando linguagem natural ou palavras-chave em uma linguagem L1 e obtenha resultados na linguagem L2. De tal modo, pode-se, por exemplo, realizar consultas em Português obtendo resultados em outro idioma. Foi considerado as publicações dos últimos cinco anos (2015-2019) indexadas por seis bases científicas reconhecidas internacionalmente.*

1. Introdução

Com o exponencial crescimento do volume de dados armazenados em bancos de dados relacionais e a globalização da internet, a possibilidade de consultar informações em vários idiomas - consulta multilíngue[1] - se tornou uma tarefa importante para a obtenção da informação desejada. Esse tipo de consulta permite aos usuários obterem acesso à informações mais precisas e completas utilizando linguagem natural ou palavras-chave. Com isso, a Recuperação da Informação em outro idioma tem atraído a atenção dos pesquisadores, uma vez que permite expandir os resultados da consulta, além de ampliar a probabilidade de obter melhores resultados que coincidam com a intenção do usuário.

Em uma arquitetura padrão - consulta monolíngue - o usuário informa os termos da consulta inicial que são pré-processados utilizando técnicas de Processamento de

[1]Este artigo adota a concepção de consulta multilíngue como a possibilidade de se realizar a consulta inicial em um idioma e, a partir de sua tradução, obter resultados em vários idiomas.

Linguagem Natural (PLN) como *tokenization*, *stopwords*, *lemmatization/stemming*. Na expansão da consulta são utilizados dicionários de sinônimos para encontrar termos similares aos da consulta inicial. Em seguida, esses termos são mapeados em expressões SQL que são executadas no Banco de Dados e retornam os resultados da consulta inicial para o usuário. Um exemplo dessa arquitetura para consulta em bancos de dados relacionais utilizando linguagem natural ou palavras-chave pode ser visualizada na figura 1.

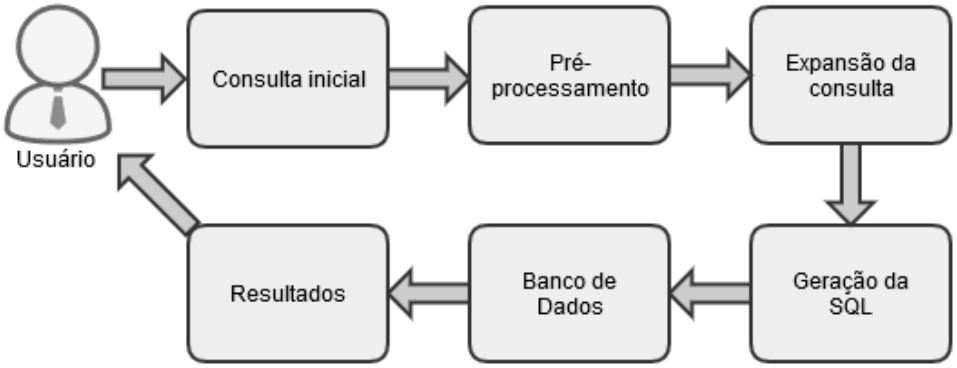

Figura 1. Arquitetura padrão

Já no caso de uma abordagem multilíngue a tradução da consulta inicial pode ser realizada antes ou depois do pré-processamento. Quando realizada antes a tradução é realizada na sentença completa preservando uma maior semântica na tradução e aumentando as entradas para o pré-processamento. Se realizada após o pré-processamento, na expansão da consulta, a tradução é realizada termo a termo podendo aumentar a ambiguidade e retornar resultados inconsistentes.

Existem diferentes métodos para realizar uma consulta multilíngue. Entre eles há o *Cross-Language Information Retrieval (CLIR)*, um sistema que aceita consultas em um determinado idioma para consultar dados em qualquer idioma diferente do idioma de origem [Poibeau 2017], por exemplo, uma consulta no idioma português com resultados em inglês. Por meio dele é possível realizar a tradução da consulta inicial para obter resultados em outro idioma conforme foi utilizado nas propostas de [Chandra and Dwivedi 2019], [Thenmozhi and Aravindan 2018], [Valiveti et al. 2017], [Rahmani 2017] e [Singla et al. 2016]. Assim, conforme [Madankar et al. 2016], o CLIR trata as consultas em um idioma e a recuperação de resultados em outro idioma.

Outro método que pode ser utilizado é o *Multilanguage Information Retrieval (MLIR)* que possibilita submeter uma consulta ou obter resultados em vários idiomas por meio da tradução, por exemplo, a consulta pode ser realizada em inglês ou francês e os resultados em português e inglês. Esse método é abordado em [Satyamurty et al. 2018],[Virk and Dua 2017] e [Posevkin and Bessmertny 2016]. De acordo com [Madankar et al. 2016], o MLIR trata consultas em um ou mais idiomas e a recuperação de resultados também em um ou mais idiomas.

A principal diferença entre o CLIR e o MLIR é que no CLIR há restrição de um único idioma para consulta inicial ou obtenção de resultados sendo que o idioma da consulta deve ser diferente do idioma dos resultados. Já no MLIR a consulta pode ser realizada em vários idiomas, bem como os resultados obtidos. Esses métodos podem realizar

a consulta multilíngue por meio de diferentes tipos de tradução, tais como: da consulta inicial; do resultado da consulta; ou da consulta inicial e do resultado da consulta. O método de tradução da consulta inicial tem sido o mais utilizado devido a sua simplicidade e, conforme [Rahmani 2017], tem atraído muita atenção devido seu desempenho.

No caso da tradução da consulta inicial podem ser utilizados três meios principais:

- **Dicionário:** uso de um dicionário bilíngue para traduzir de uma linguagem L1 para a linguagem L2;
- **Tradução Automática:** uso de um software para traduzir o resultado ou a consulta de uma linguagem L1 para outras linguagens;
- **Corpora Paralelo:** uso de um corpo de documentos em duas ou mais línguas como referência para realizar novas traduções.

Com base nesses apontamentos, a principal motivação deste estudo é analisar as abordagens propostas para realizar consultas multilíngues em banco de dados relacionais identificando as limitações, lacunas, problemas encontrados e soluções propostas de forma a sintetizar o conhecimento na área e proporcionar novas abordagens. A problemática reside no fato de que os métodos de busca tradicionais são monolíngues o que limita os resultados para o idioma da consulta inicial. Assim, para obter resultados em outros idiomas é necessário reformular a consulta para o idioma pretendido. Essa nova consulta pode retornar resultados indesejados se for utilizado termos inadequados devido a falta de domínio na nova língua.

O objetivo geral desta pesquisa é realizar uma revisão sistemática dos principais estudos relacionados a temática de consulta multilíngue em bancos de dados relacionais. Tem como objetivos específicos descrever quais são os métodos utilizados para realizar a etapa de pré-processamento dos termos da consulta e identificar os métodos para realizar a tradução dos termos da consulta inicial em sistemas de consulta multilíngue.

Assim, este artigo apresenta os principais métodos de pré-processamento e tradução da consulta em bancos de dados relacionais. Para tanto, baseia-se na análise das publicações científicas acerca da temática em questão - dos últimos cinco anos (2015-2019) - indexadas em reconhecidas bases científicas. Para além da introdução, o trabalho aborda a metodologia de pesquisa planejada para realizar a revisão sistemática; explica os passos executados na condução dessa revisão; e apresenta as conclusões sobre a análise dos principais trabalhos encontrados.

2. Metodologia de pesquisa

O objetivo dessa revisão sistemática é apresentar o estado da arte de estudos relacionados aos mecanismos de consulta multilíngue em bancos de dados relacionais. O banco de dados relacional é o principal meio de armazenamento de dados das organizações, portanto foi escolhido como objeto de estudo por conter uma maior representatividade. Os estudos apresentados enfatizam a utilização da tradução da consulta inicial no retorno de resultados em outros idiomas como forma de ampliar o acesso e a relevância das informações disponibilizadas.

O protocolo da revisão sistemática foi dividido nas seguintes etapas: definição das questões de pesquisa, das palavras-chave, dos sinônimos e estratégia da pesquisa, dos critérios de inclusão e exclusão dos trabalhos, da avaliação da qualidade dos trabalhos

selecionados e da extração de dados por meio de um formulário de perguntas. Cada um desses elementos será detalhado no decorrer desse estudo. Ademais, foi utilizada a ferramenta Parsifal [2] para apoiar a implementação da metodologia proposta.

2.1. Questões de pesquisa

O presente estudo é baseado nas seguintes questões de pesquisa: Quais os principais métodos para tradução da consulta inicial podem ser utilizados em bancos de dados relacionais? Quais métodos e técnicas são utilizados no pré-processamento da consulta inicial para realizar a tradução?

2.2. Palavras-chave, sinônimos e estratégia

Para identificar os principais estudos relacionados com a temática foi utilizada a seguinte expressão genérica de busca (contendo palavras-chave e seus sinônimos): *(('natural language' OR keyword) AND (query OR search OR 'information retrieval') AND (multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual) AND ('relational database' or database))*.

A *string* de busca foi elaborada utilizando os termos recorrentes de trabalhos da área *keyword search over relational databases* e os termos foram ajustados para ampliar a busca nas áreas de linguagem natural e consultas multilíngues. A *string* gerada foi executada nas principais bases de dados científicas da área: ACM Digital Library [3], IEEExplore [4], Science Direct [5], Scopus [6], Spring Link [7] e Web of Science [8]. Para escolha das bases de dados foi utilizado como critério a possibilidade de exportar os resultados para o formato *.bibtex* nativamente ou utilizando ferramentas externas. Ademais foram eliminadas as bases que não retornaram resultados ou que os resultados foram idênticos aos de outras bases já selecionadas. Os resultados foram filtrados para o período compreendido entre os anos de 2015 a 2019.

2.3. Critérios de inclusão/exclusão

Foram definidos critérios de inclusão e exclusão para serem utilizados como referência na seleção dos resultados obtidos com a *string* de busca. Nesta etapa, foi analisado o título, o resumo e as palavras-chave dos trabalhos para incluir/aceitar ou para excluir/rejeitar os trabalhos.

Critérios de inclusão:

1. Obtido no resultado da string de busca;
2. Acesso disponível nas bases científicas selecionadas;
3. Propõe tradução da consulta inicial.

Critérios de exclusão:

1. O estudo não possui conteúdo relevante para o objetivo da pesquisa;

[2]https://parsif.al
[3]https://www.dl.acm.org
[4]https://ieeexplore.ieee.org
[5]https://www.sciencedirect.com
[6]https://scopus.com
[7]https://link.springer.com
[8]https://apps.webofknowledge.com

2. Artigo sem acesso público ou institucional;
3. Publicado como pôster ou resumo;
4. Publicado antes de 2015.

2.4. Avaliação da qualidade

Para avaliar a qualidade dos trabalhos selecionados com a aplicação dos critérios de inclusão/exclusão foram definidas as seguintes perguntas (cada uma podendo apresentar como resposta: sim/talvez/não) e a respectiva pontuação (10/3/0):

1. Permite consulta multilíngue?
2. Data de publicação maior ou igual a 2017?
3. Apresenta conteúdo relevante?
4. Apresenta proposta similar?

Esses critérios foram definidos para priorizar a ordem de leitura dos artigos que obtiveram uma maior pontuação de qualidade. São considerados mais relevantes os artigos que abordam a consulta multilíngue publicados recentemente (a partir de 2017) e que apresentam conteúdo relevante ou proposta similar à realização de consulta multilíngue em bancos de dados relacionais.

2.5. Formulário de extração de dados

Foi elaborado um formulário para realizar a extração de dados contendo as seguintes perguntas:

1. Quais técnicas de pré-processamento foram usadas?
2. Quais técnicas de tradução foram usadas?
3. Possui independência de domínio?
4. Possui independência de banco de dados?
5. Necessita de seleção do idioma?
6. Permite apenas instruções de consulta?
7. Retorna resultados em vários idiomas?
8. Permite consultar em mais de um idioma?
9. Possui recurso de autocompletar a consulta?
10. Usa banco de dados relacional?

3. Condução da revisão

A etapa de condução foi responsável por efetivamente realizar o processo da revisão sistemática. O processo foi planejado executando as *strings* de busca nas bases de dados científicas selecionadas, exportando os trabalhos resultantes para o formato .bibtex e importando na ferramenta utilizada para gerenciar a revisão sistemática. Em seguida, foi realizada a seleção dos estudos aplicando os critérios de inclusão e exclusão. Dentre os trabalhos selecionados foi realizada a avaliação da qualidade para então realizar a extração de dados.

3.1. *Strings* de busca

Devido às especifidades de cada base científica a *string* de busca genérica foi adaptada para cada uma delas e utilizada no campo de pesquisa avançada da seguinte forma:

- **ACM (Título e Resumo):** +('natural language' keyword) +(query search 'information retrieval') +(multilingual multilanguage cross-language cross-lingual bilingual) +('relational database' database);
- **IEEE (Texto completo e Metadados):** ('natural language' OR keyword) AND (query OR search OR 'information retrieval') AND (multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual) AND ('relational database' or database);
- **Science Direct:** ('natural language' OR keyword) (query OR search OR 'information retrieval') (multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual) ('relational database' or database);
- **Scopus:** (TITLE-ABS-KEY('natural language' OR keyword) OR TITLE-ABS-KEY(query OR search OR 'information retrieval') AND TITLE-ABS-KEY(multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual) AND TITLE-ABS-KEY('relational database' or database));
- **Springer Link:** ('natural language' OR keyword) AND (query OR search OR 'information retrieval') AND (multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual) AND ('relational database' or database);
- **Web of Science:** ALL=('natural language' OR keyword) AND ALL=(query OR search OR 'information retrieval') AND ALL=(multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual) AND ALL=('relational database' or 'database').

Todos os resultados foram exportados para o formato .bibtex e importados na ferramenta Parsifal que facilita o processo de revisão sistemática. Para a base Springer Link foi utilizada a ferramenta JabRef [9] para importar os documentos resultantes da *string* de busca e gerar o arquivo .bibtex, pois esta base não possui a funcionalidade de exportar.

3.2. Importação dos estudos

Foram importados 833 estudos distribuídos por base de dados científica conforme podemos observar na figura 1:

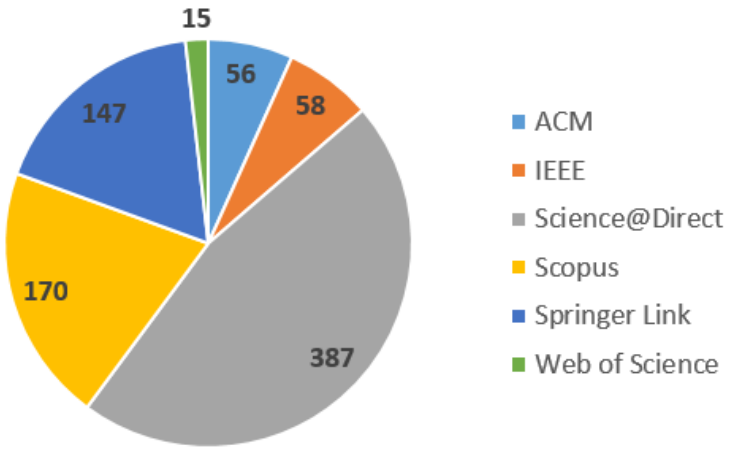

Figura 2. Estudos importados

[9]http://www.jabref.org/

3.3. Seleção dos estudos

A seleção dos estudos foi realizada com a leitura do título, resumo e palavras-chave de cada documento, observando se eles se enquadravam em algum critério de inclusão ou exclusão. Dentre os 833 trabalhos importados das bases científicas, 31 trabalhos foram classificados como duplicados e portanto não foram analisados, 792 foram rejeitados e 10 trabalhos foram aceitos utilizando os critérios definidos na subseção 2.3.

3.4. Avaliação da qualidade

Para avaliação da qualidade foi analisado o título, resumo, palavras-chave e, quando necessário, a introdução do documento para identificar as possíveis respostas para os questionamentos que foram definidos na etapa de planejamento da avaliação da qualidade. A Tabela 1 apresenta a pontuação de cada pergunta e a pontuação total de qualidade obtida para cada artigo aceito, conforme definido na subseção 2.4.

Tabela 1. Avaliação da qualidade das publicações aceitas

Referência	Q1	Q2	Q3	Q4	Total
[Satyamurty et al. 2018]	10	10	10	10	40
[Thenmozhi and Aravindan 2018]	10	10	10	3	33
[Chandra and Dwivedi 2019]	10	10	3	3	26
[Virk and Dua 2017]	10	10	3	3	26
[Singla et al. 2016]	10	0	10	3	23
[Choudhary et al. 2015]	10	0	10	3	23
[Posevkin and Bessmertny 2016]	10	0	10	3	23
[Madankar et al. 2016]	10	0	10	0	20
[Valiveti et al. 2017]	3	10	0	3	16
[Rahmani 2017]	3	10	0	0	13

Na Tabela 1, podemos identificar que dentre as dez publicações aceitas somente uma obteve a pontuação máxima de qualidade atendendo a todos os critérios de qualidade selecionados. Isso evidencia que esse ainda é um problema em aberto da área, pois há poucas propostas que abordam a interseção dos temas: consultas por palavra-chave ou linguagem natural, banco de dados relacionas e consulta multilíngue.

3.5. Extração de dados

Para a extração dos dados foi realizada a leitura completa dos trabalhos priorizando aqueles que obtiveram maior pontuação na etapa de avaliação da qualidade da subseção 2.4. Em seguida, foi realizada a classificação dos trabalhos de acordo com as questões do formulário de extração de dados que foram definidas na subseção 2.5. O resultado dessa classificação é apresentado na Tabela 2.

3.6. Análise dos resultados

Diante da análise realizada ficou evidente que poucos trabalhos descrevem o processo ou arquitetura de pré-processamento e de tradução da consulta inicial. Assim, a análise revelou que o uso de técnicas de pré-processamento de linguagem natural, tais como *tokenization, stemming ou lemmatization,* e remoção de *stopwords*, foi observado nos

Tabela 2. Extração de dados das publicações aceitas

Referência	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
[Satyamurty et al. 2018]	-	Microsoft Translate	Não	Não	Não	Sim	Sim	Não	Não	Sim
[Thenmozhi and Aravindan 2018]	Lemma, Stopwords, POS e Token	Dicionário bilíngue	Não	Não	Não	Sim	Não	Não	Não	Não
[Chandra and Dwivedi 2019]	-	Google Translate	Sim	Sim	Não	Sim	Sim	Não	Não	Não
[Virk and Dua 2017]	Léxica e sintática	-	Sim	Não	Sim	Sim	Não	Sim	Sim	Sim
[Singla et al. 2016]	Stopwords, NER, Lemma e Token	-	Sim	Não	Sim	Sim	Não	Sim	Não	Não
[Choudhary et al. 2015]	Token, POS e Chunking	-	Não	Não	Sim	Sim	Não	Sim	Sim	Sim
[Posevkin and Bessmertny 2016]	-	-	Não	Não	Não	Sim	Não	Não	Não	Sim
[Madankar et al. 2016]	-	-	Não	Não	Não	Sim	Não	Não	Não	Sim
[Valiveti et al. 2017]	-	-	Não	Não	Não	Sim	Sim	Não	Não	Sim
[Rahmani 2017]	-	Google Translate	Não	Não	Não	Sim	Não	Não	Não	Não

trabalhos de [Thenmozhi and Aravindan 2018], [Virk and Dua 2017], [Singla et al. 2016] e [Choudhary et al. 2015].

A tradução da consulta que é a parte essencial para consultas multilíngues foi citado apenas nos trabalhos de [Chandra and Dwivedi 2019], [Thenmozhi and Aravindan 2018], [Satyamurty et al. 2018] e [Rahmani 2017]. Ana-

lisando esses trabalhos ficou notório que o *Google Translate*[10] foi a ferramenta que prevaleceu no processo de tradução utilizado por eles (dentre os quatro trabalhos dois utilizaram essa ferramenta). Provavelmente, isso se deve ao fato de ser uma ferramenta globalizada, amplamente difundida e com suporte para mais de 100 idiomas, além de estar em constante evolução apresentando resultados de qualidade.

A técnica de expansão de consulta[11] foi abordada no trabalho de [Chandra and Dwivedi 2019] no qual foram realizadas 50 consultas no idioma Hindi que foram traduzidas para o idioma Inglês utilizando o *Google translate*. Os termos em Inglês obtidos foram expandidos para obter resultados mais relevantes. Essa técnica também foi apresentada no trabalho de [Thenmozhi and Aravindan 2018]. A expansão da consulta foi realizada usando a *WordNet* para reformular a consulta inicial e reduzir as ambiguidades dos termos da consulta após a tradução para outro idioma. Nesse trabalho também foi descrito um analisador morfológico que identifica a forma raíz dos *tokens* extraídos, bem como um método de tradução que utiliza um dicionário bilíngue contendo os principais termos do domínio de agricultura.

Na arquitetura proposta por [Choudhary et al. 2015] a *WordNet* foi utilizada como um dicionário de sinônimos no módulo de reconhecimento de domínio para ampliar os resultados na identificação do domínio e responder à consulta facilmente. A *WordNet* também foi utilizada no modelo proposto por [Satyamurty et al. 2018] que substitui as palavras-chave da consulta inicial por palavras equivalentes semanticamente no idioma de destino.

A interface em linguagem natural para banco de dados que realiza a conversão de linguagens indianas como *Gujarati* para o idioma Inglês foi proposto no trabalho de [Valiveti et al. 2017]. Esse trabalho menciona um algoritmo de tolerância na seção de revisão de literatura que obtém como entrada uma *string* na língua local, extrai as vogais e retorna o resultado na língua Inglesa. No entanto, o estudo não descreve as técnicas de pré-processamento e não menciona o método utilizado para realizar a tradução dos termos para outras línguas.

O uso de um *framework* para traduzir as palavras-chave do usuário numa linguagem de destino usando o *Microsoft translate* foi proposto no trabalho de [Satyamurty et al. 2018] que também utilizou a WordNet para realizar a expansão da consulta no domínio educacional. Nos experimentos foram utilizados os idiomas Espanhol e Japonês como idioma de destino. Já no trabalho de [Rahmani 2017] foram apresentadas as métricas BLEU e MAP para avaliar as traduções realizadas usando o *Google translate* a fim de traduzir do Inglês para a língua Persa no domínio médico. O autor propõe uma nova abordagem para sistemas *cross-lingual information retrieval*, mas não detalha os métodos ou técnicas que foram utilizados.

O desenvolvimento de um sistema independente de domínio para banco de dados nas linguagens *Punjabi* e *Hindi* com recurso de completar automaticamente foi proposto em [Virk and Dua 2017]. O banco de dados utilizado possui valores de dados nas duas línguas indianas e os metadados estão no idioma Inglês. Nesse estudo, a linguagem utili-

[10] `https://translate.google.com/intl/en_ALL/about/languages`

[11] Essa técnica ajuda a melhorar a qualidade e relevância dos resultados recuperados, pois permite expandir a consulta para os sinônimos dos termos da consulta

zada na consulta era informada manualmente e havia um módulo para identificar o banco de dados de acordo com os *tokens*, palavras-chave e nomes de campo da consulta inicial.

Já no trabalho de [Singla et al. 2016], foi projetada uma arquitetura bilíngue na qual o usuário realiza a escolha do idioma da consulta que seria pré-processada gerando *tokens* e removendo as *stopwords*. Em seguida, se realizava a identificação da entidade utilizando o algoritmo *Smith Waterman* para calcular a similaridade entre dois objetos e finalizar classificando os resultados.

[Choudhary et al. 2015] propõe em seu estudo um sistema bilíngue que aceita as línguas *Hindi* e *Punjabi* na consulta e obtém os resultados na mesma língua pesquisada. Na proposta as consultas executadas com sucesso são armazenadas para sugerirem novas consultas automaticamente.

[Posevkin and Bessmertny 2016] apresenta uma arquitetura que identifica o idioma da consulta, extrai os *tokens*, a classe gramatical, realiza análise morfológica, remove as *stopwords* e identifica o domínio da consulta. Além disso, propõe um protótipo que possibilita consultar um banco de dados com metadados em Inglês utilizando a distância de *Levenshtein* e processo de decisão de *Markov*.

O recurso de autocompletar a consulta inicial - que auxilia o usuário no processo de pesquisa, possibilita a redução de erros ortográficos e sugere expressões já obtidas em resultados anteriores - foi observado nos trabalhos de [Virk and Dua 2017] e [Choudhary et al. 2015]. Já o trabalho de [Madankar et al. 2016] apresenta uma revisão sobre sistemas de recuperação da informação e tradução automática e descreve os principais conceitos de recuperação da informação multilíngue.

No estudo proposto por [Rahmani 2017] foi utilizada a métrica BLEU [Papineni et al. 2002] para medir a precisão da tradução automática em comparação com a tradução feita pelo ser humano. Vale ressaltar que para avaliar o desempenho dos sistemas de tradução automática podem ser utilizadas, para além da métrica BLEU, as métricas NITS [Doddington 2002] e METEOR [Banerjee and Lavie 2005].

Com base nos apontamentos descritos fica evidente que os estudos analisados realizam somente a operação de consulta de dados, apesar de existir possibilidade de adaptação das técnicas para realizarem operações de inserção, atualização e remoção. Ademais, a maioria dos estudos analisados são dependentes de domínio e de banco de dados, mostrando que os resultados foram obtidos em um ambiente restrito e controlado. Metade dos trabalhos analisados (cinco) apresentaram propostas voltadas para a consulta em banco de dados relacionais. Por fim, dentre os estudos analisados, os trabalhos de [Virk and Dua 2017, Singla et al. 2016, Choudhary et al. 2015] demonstram que é possível informar a consulta inicial em mais de um idioma mediante seleção manual restringindo os resultados da consulta para o idioma selecionado.

4. Conclusões

Com a análise dos estudos encontrados sobre consulta multilíngue em banco de dados relacionais foi possível observar que é uma área que está começando a despertar o interesse dos pesquisadores. Isso devido a constante evolução das principais técnicas utilizadas no Processamento de Linguagem Natural para realizar o pré-processamento e tradução de consultas. Essa situação tem despertado a atenção de pesquisadores de todo o mundo que

buscam romper as barreiras linguísticas e ampliar o acesso à informação ao possibilitar consultas multilíngues.

A análise revelou que mais da metade dos estudos envolvem propostas de consultas multilíngues com abordagens diferentes, tais como: realizar a tradução da consulta inicial para obter resultados em um idioma universal como o inglês; informar a consulta em vários idiomas para obter resultados em vários idiomas consultados; e traduzir os resultados obtidos para o idioma utilizado na consulta inicial.

A revisão realizada aponta ainda que há vários problemas a serem tratados, principalmente os relacionados as variações linguísticas e as particularidades de cada idioma. Apesar disso, já é possível observar melhorias significativas relacionadas a essas questões. O crescente aumento da capacidade dos recursos computacionais e a evolução das técnicas de tradução tem contribuído para aumentar a efetividade da tradução e fomentar as mais diversas aplicações de tradução em sistemas de busca multilíngues.

Ademais, com base nos resultados da revisão ficou evidente que ainda é necessário ampliar os estudos sobre a identificação automática do idioma da consulta. Isso com a finalidade de encontrar uma forma de dispensar a seleção manual pelo usuário evitando possíveis inconsistências decorrentes de uma seleção incorreta. Assim, um melhor detalhamento das abordagens propostas, dos métodos de pré-processamento e da tradução da consulta são importantes para possibilitar evoluções das abordagens propostas pela comunidade científica.

Vale ressaltar que as principais limitações dos trabalhos encontrados na revisão sistemática realizada nesse artigo consistem em: dependência de domínio que dificulta a generalização da abordagem para outros contextos e dependência do banco de dados que restringe a solução para banco de dados específicos.

Referências

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chandra, G. and Dwivedi, S. K. (2019). Query expansion for effective retrieval results of hindi–english cross-lingual IR. *Applied Artificial Intelligence*, 33(7):567–593.

Choudhary, M., Dua, M., and Virk, Z. S. (2015). A web-based bilingual natural language interface to database. In *2015 Third International Conference on Image Information Processing (ICIIP)*. IEEE.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Madankar, M., Chandak, M., and Chavhan, N. (2016). Information retrieval system and machine translation: A review. *Procedia Computer Science*, 78:845–850.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Poibeau, T. (2017). *Machine Translation*. The MIT Press Essential Knowledge Series, London, England, 1 edition.

Posevkin, R. and Bessmertny, I. (2016). Multilanguage natural user interface to database. In *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE.

Rahmani, A. (2017). Adapting google translate for english-persian cross-lingual information retrieval in medical domain. In *2017 Artificial Intelligence and Signal Processing Conference (AISP)*. IEEE.

Satyamurty, C. V. S., Murthy, J. V. R., and Raghava, M. (2018). Metadata-based semantic query in multilingual databases. In *Advances in Intelligent Systems and Computing*, pages 249–255. Springer Singapore.

Singla, K., Dua, M., and Nanda, G. (2016). A language based comparison of different similarity functions and classifiers using web based bilingual question answering system developed using machine learning approach. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies - ICTCS 16*. ACM Press.

Thenmozhi, D. and Aravindan, C. (2018). Ontology-based tamil-english cross-lingual information retrieval system. *Sādhanā*, 43(10).

Valiveti, S., Tripathi, K., and Raval, G. (2017). Natural language interface for multilingual database. In *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2*, pages 113–120. Springer International Publishing.

Virk, Z. and Dua, M. (2017). An advanced web-based bilingual domain independent interface to database using machine learning approach. In *Advances in Intelligent Systems and Computing*, pages 581–589. Springer Singapore.