

RECUPERAÇÃO DE DADOS EM BANCO DE DADOS POR MEIO DA LINGUAGEM NATURAL

DATA RETRIEVAL IN RELATIONAL DATABASES USING NATURAL LANGUAGE

Bruno Perché Pinto¹, Denis Pinheiro², Denilson Alves Pereira³

^{1,2}Centro Universitário de Belo Horizonte – Belo Horizonte, MG; ³Departamento de Ciência da Computação - Universidade Federal de Lavras – Lavras, MG

¹brunoperche@gmail.com, ²denispinheiro@gmail.com, ³denilsonpereira@dcc.ufla.br

Resumo – A Linguagem Natural é uma alternativa para consultas em bancos de dados, principalmente para usuários leigos em computação, que provavelmente têm dificuldades em utilizar linguagens de consulta como o SQL ou similares. Este trabalho apresenta um protótipo de uma ferramenta para a recuperação de dados em bancos de dados utilizando Linguagem Natural em Português. Serão apresentadas algumas técnicas e métodos para o Processamento de Linguagem Natural (PLN), mostrando suas principais aplicações e importância para recuperar a informação em banco de dados.

Palavras-chave – Recuperação de Informação. Processamento de Linguagem Natural. Banco de Dados. Inteligência Artificial.

Abstract – A Natural Language is an alternative to queries in databases, especially for layman users, which probably have difficulty using query languages like SQL or similar. This paper presents a prototype of a tool for data retrieval in databases using Natural Language in Portuguese. There will be presented some techniques and methods for Natural Language Processing (NLP), showing its main applications and importance to retrieve the information in the database.

Keywords – Information Retrieval. Natural Language Processing. Database. Artificial Intelligence.

I INTRODUÇÃO

O homem tem armazenado, catalogado e organizado a informação há aproximadamente

4.000 anos, com o principal objetivo de recuperá-la para uso posterior. Atualmente, cresce de forma cada vez mais rápida a quantidade dos textos armazenados em formato digital, e a maioria destes textos são esquecidos, pois nenhum ser humano pode ler, entender e sintetizar toda esta informação. Isto tem incentivado os pesquisadores a explorar estratégias para tornar acessível ao usuário a informação relevante (CROFT, 2000).

A obtenção da informação através de consultas constitui a entrada de um sistema de recuperação de informação, que deve buscar em uma coleção de documentos, aqueles que podem satisfazer à necessidade de informação do usuário. Entretanto, é difícil encontrar a informação relevante, principalmente porque há muita informação (a maioria irrelevante). O Processamento da Linguagem Natural (PLN), com motivação linguística, é uma das alternativas para solucionar este problema (JACQUEMIN, 2000).

Muitos pesquisadores acreditam que o PLN seja uma das áreas mais importantes da Inteligência Artificial. Uma de suas aplicações mais atraentes é a possibilidade de tornar o relacionamento do usuário com a máquina o mais natural possível, facilitando o seu acesso às

informações que este só poderia obter por meio de conhecimentos de alguma linguagem de programação, ou por meio de alguém que tenha esse conhecimento. Seria muito mais fácil se os usuários pudessem se comunicar com o computador por meio de conversação. O objetivo do PLN é tornar isso uma realidade (MURADAS, 1995).

Um dos grandes problemas para usuários leigos, relacionados à recuperação de dados em bancos de dados, está ligado à interface e à necessidade de um pré-conhecimento da linguagem estruturada de consulta: SQL (Structured Query Language).

No contexto de bancos de dados relacionais, embora SQL seja uma linguagem de aprendizado relativamente fácil, pessoas que não possuem formação na área de computação podem encontrar certa dificuldade em aprendê-la. Além do conhecimento relacionado à linguagem SQL, os usuários precisam saber os conceitos envolvidos com o esquema da base de dados, o que implica em ter um mínimo de conhecimento do modelo relacional lógico e físico (tabelas, chaves primárias e estrangeiras, junções, implementação de relacionamentos $m \times n$). Sendo assim, usuários leigos, não familiarizados com linguagens de consulta e estruturas de bancos de dados, deveriam poder acessar os dados de uma forma mais amigável (VARELLA et al., 2006).

Uma forma de facilitar o uso de consultas em banco de dados é através de técnicas de PLN, onde usuários podem extrair informações sem a necessidade de conhecer particularidades de um modelo de dados ou de estruturas de armazenamento. Por exemplo, em um banco de dados acadêmico, o usuário poderia formular a seguinte consulta:

“Quais são os nomes dos alunos matriculados e os cursos que estudam?”

Em SQL o usuário teria que formular o seguinte comando:

“select aluno.nome, curso.nome from aluno, curso where aluno.idcurso = curso.idcurso”.

Nota-se que para a formulação da consulta em PLN, não é necessário que o usuário saiba detalhes sobre a estrutura do banco de dados (meta-informação); já em SQL essas informações são cruciais, inclusive a notação exata dos termos no banco de dados (campo *nome do aluno* da tabela *aluno* e *nome do curso* da tabela *curso*), além da construção da junção relacional (*join*) e da sintaxe rígida da linguagem SQL (SILVA; LIMA, 2007).

O presente trabalho tem como objetivo tratar de um problema de grande importância prática, a implementação de uma ferramenta que através de técnicas de Processamento de Linguagem Natural permita entender sentenças em português que especifique uma consulta em uma base de dados e transformá-las em consultas SQL, retornando assim as devidas informações solicitadas.

Trata-se de uma pesquisa exploratória, dado que PLN em língua portuguesa é pouco estudado. O desenvolvimento das ferramentas e a aplicação para essa área reforçam a classificação exploratória.

Mais de um critério pode ser usado na avaliação de uma ferramenta de criação de consultas. O foco deste protótipo também está relacionado à usabilidade. Assim, algumas funções não foram implementadas, como funções agregadas, cláusulas *group by*, *order by*, *having*, *distinct* e as operações que alteram o estado de banco de dados.

A ferramenta desenvolvida em (SOUZA et al., 2006) tem como objetivo permitir que usuários leigos construam suas consultas utilizando uma linguagem próxima à linguagem natural. O usuário é conduzido passo a passo, na formulação da consulta até que a defina completamente. A cada passo, a ferramenta apresenta um conjunto de palavras ou frases,

uma das quais o usuário escolhe como parte da consulta. O processo é interativo e prossegue até que a mesma esteja completa e seja submetida para execução.

O trabalho desenvolvido em (SILVA; LIMA, 2007) apresenta técnicas de PLN para consultas em banco de dados com a implementação do analisador sintático na linguagem de programação Prolog. Originalmente esta linguagem de programação foi projetada para o PLN, possuindo algumas características vantajosas, dentre elas um formalismo para representação de gramáticas livres de contexto denominado DCG (Definite Clause Grammar), associado a um analisador top-down recursivo, da esquerda para a direita.

Para análise das palavras em linguagem natural, processo conhecido como etiquetagem das palavras, foi utilizado o módulo desenvolvido em (NUNES, 2007).

O presente trabalho está organizado nas seguintes seções:

A seção 1 apresenta a introdução.

Na seção 2 consta o referencial teórico mencionando a Recuperação da Informação, Inteligência Artificial e Processamento de Linguagem Natural.

A seção 3 aborda o desenvolvimento, detalhando os procedimentos realizados.

Na seção 4 são apresentadas as implementações que tratam a arquitetura da solução e os detalhes da implementação.

Na seção 5 constam as análises e resultados obtidos neste projeto.

A seção 6 aborda as possíveis linhas de pesquisa para trabalhos futuros.

Na seção 7 são apresentadas as conclusões que puderam ser identificadas com este trabalho.

II REFERENCIAL TEÓRICO

O referencial teórico deste artigo foi construído com base nas temáticas: Recuperação da Informação, Inteligência Artificial, Processamento de Linguagem Natural e XML. A seguir encontra-se uma sucinta introdução que evidencia a importância de cada um desses ramos de conhecimento para o estudo em questão.

II.1 RECUPERAÇÃO DA INFORMAÇÃO

A Recuperação da Informação é um tema investigado tanto pela Ciência da Informação quanto pela Ciência da Computação. A Recuperação da Informação tem sido estudada pela Ciência da Informação desde a sua origem, quando essa área apontou como sua finalidade a tarefa de tornar acessível um acervo crescente de registros de conhecimentos (DUQUE, 2005).

Segundo Cardoso (2007), a Recuperação da Informação é uma área da Ciência da Computação que lida com armazenamento automático e recuperação de documentos, que são de grande importância devido ao uso universal da linguagem para comunicação.

Grande parte das informações armazenadas pelo homem se encontra, atualmente, em formato digital. Uma das preocupações dos pesquisadores da área da Recuperação da Informação está em como organizar e representar as inserções de novos documentos neste acervo, bem como facilitar a busca aos documentos, em uma posterior recuperação (MIORELLI, 2001).

Três conceitos fundamentais são utilizados para avaliação de sistemas de Recuperação de Informação. O primeiro deles é o conceito de *relevância*. Um documento relevante é aquele que atende a necessidade do usuário, a qual foi submetida como uma pesquisa à base.



O conceito de *revocação* refere-se à medida que avalia o quanto foi possível abranger os documentos relevantes da base de dados em uma pesquisa.

Por fim, o conceito de *precisão* associa-se à medida que avalia o quanto o resultado da consulta foi efetivamente útil (CÂMARA JÚNIOR, 2007).

Os conceitos abordados anteriormente relativos à relevância, revocação e precisão são demonstrados na Figura 1.



Figura 1- Definição de Relevância, Revocação e Precisão.
Fonte: (CÂMARA JÚNIOR, 2007).

A Figura 1 demonstra o quanto o resultado da pesquisa foi efetivo no aspecto de evitar que documentos irrelevantes poluam o retorno.

II.2 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial é a parte da Ciência da Computação que estuda o desenvolvimento de sistemas computacionais com características que podem ser associadas à inteligência no comportamento humano, tais como: compreensão da linguagem, aprendizado, raciocínio, resolução de problemas, etc. (FERNANDES, 2003).

Segundo Vinhaes (2005), a Inteligência Artificial é uma ciência nova, iniciada na década de quarenta, baseada em pesquisas que estudavam o funcionamento do cérebro com o

objetivo de formalizar seu comportamento. Mas somente a partir de 1956, a Inteligência Artificial foi reconhecida como uma ciência. Ela se baseou em várias áreas do conhecimento humano, tais como: neurociência (que fornece conhecimentos sobre o funcionamento do nosso cérebro), filosofia (estuda como é o pensamento humano), entre outras. Entre estas e outras áreas, a linguística é essencial para o desenvolvimento da PLN, pois além de fornecer formas de representar o conhecimento, ela propicia todos os conhecimentos necessários para o desenvolvimento de um sistema que seja capaz de processar alguma Linguagem Natural.

Em 1950, Alan Turing, propôs um teste para fornecer uma definição operacional do que seria inteligência. Para Turing, uma máquina poderia ser considerada inteligente se ela fosse capaz de responder perguntas feitas por um interrogador, e que o mesmo não saberia se as respostas estavam vindo de uma pessoa ou não. Uma máquina para passar neste teste teria que ter as seguintes capacidades:

- PLN: capacidade de comunicação com um idioma natural;
- Representação do conhecimento: capacidade de armazenar conhecimento de tudo o que foi aprendido;
- Raciocínio automatizado: capacidade de utilizar todo conhecimento adquirido, e tirar novas conclusões;
- Aprendizado de máquina: poder adaptar-se a novas circunstâncias (RUSSEL; NORVIG, 2004).

II.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Na década de 60, surgiram várias aplicações com compreensão da Linguagem Natural. Nesta época, já existiam aplicações capazes de aceitar e de responder a questões em inglês, que simulavam entrevistas psiquiátricas e várias outras aplicações que auxiliavam nas mais

diferentes áreas: álgebra, medicina, relações de parentesco, etc. (OLIVEIRA, 2004).

O programa que mais se destacou nesta época foi o **Eliza**, desenvolvido por Joseph Weizenbaum, em 1966. Este programa simulava ser um psiquiatra onde o usuário interagia com a aplicação escrevendo sobre seus problemas, todavia é importante mencionar que o programa não tinha compreensão do tema que estava sendo abordado, ele apenas utilizava truques semânticos (OLIVEIRA, 2004).

Vinhaes (2005) define a Linguagem Natural como todo sistema do qual é possível extrair significado capaz de estabelecer a comunicação entre os diversos sistemas existentes, por exemplo, comunicação entre humanos ou não, naturais ou artificiais, verbais ou não verbais.

De acordo com Specia (2000, citado por VINHAES, 2005, p. 21) o PLN, visto como uma área da Inteligência Artificial e da Linguística, surgiu com o propósito de estudar a Linguagem Natural, com o objetivo de usá-la como um meio de fazer a comunicação entre os homens e os computadores. Do ponto de vista da Ciência da Computação, as principais metas do PLN é construir sistemas computacionais que facilite esta comunicação de forma efetiva, via Linguagem Natural, através dos conhecimentos linguísticos implementados em aplicações computacionais.

O PLN é uma área de pesquisa que procura construir mecanismo de tratamento e entendimento de documentos em Linguagem Natural. Essa linguagem, por possuir características informais de comunicação, tais como dependência de contextualização, morfologia e sintaxe não rigidamente definidas, entre outras, se torna muito difícil de ser tratada computacionalmente. Diversas pesquisas na área têm surgido, principalmente para língua inglesa, porém poucas em língua portuguesa (CÂMARA JÚNIOR, 2007).

O Processamento de Linguagem Natural é um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua (VINHAES, 2005).

Segundo Duque (2005), o Processamento de Linguagem Natural, pode ser entendido como um ramo da linguística que estuda a geração e recepção automática de textos, ou seja, fazer com que máquinas sejam capazes de “ler”, “escrever” e “traduzir” textos.

II.3.1 ETAPAS DO PROCESSAMENTO DE LINGUAGEM NATURAL

De acordo com Oliveira (2004) para que um sistema computacional interprete uma sentença em Linguagem Natural é necessário manter informações morfológicas, sintáticas e semânticas, armazenadas em um dicionário, juntamente com as palavras que o sistema compreende. Em seguida é realizada a fase que trata a tradução da sentença em português para o SQL. As etapas do Processamento da Linguagem Natural estão descritas na sequência.

II.3.1.1 ANÁLISE MORFOLÓGICA

O componente morfológico é aquele que se preocupa com a forma como as unidades léxicas são apresentadas. Há a utilização de um dicionário, a fim de permitir a identificação de palavras válidas na linguagem utilizada. Uma maneira otimizável para isso é que o dicionário guarde o vocabulário em sua forma canônica, mais básica e que o sistema seja capaz de derivar as palavras mais complexas de cada entrada (CÂMARA JÚNIOR, 2007).

Segundo Specia (2000, citado por VINHAES, 2005, p. 23), a Análise Morfológica tem por objetivo dividir um texto em *tokens* para que seja possível identificar os prefixos, sufixos e a formação das raízes que compõem a língua utilizada. Como resultado, haverá palavras

isoladas e descritas juntamente com seus componentes. Para que o reconhecimento destes *tokens* seja possível, é necessário que estes sejam reconhecidos pelo domínio da língua fonte. Para fazer uma busca desses *tokens* é preciso que o padrão permita gerá-lo dentro de uma estrutura de armazenamento do léxico.

Se uma palavra for substituída por outra do mesmo tipo, dentro de uma sentença gramaticalmente válida, ela ainda continuará sendo válida. Por exemplo, substantivos por substantivos, pronomes por pronomes, verbos por verbos, etc. Dentre os mesmos tipos de palavras, há vários grupos que caracterizam o comportamento dos vocábulos das linguagens. Desta forma, a morfologia trata as palavras de acordo com sua estrutura gramatical, flexão e classificação (OLIVEIRA, 2004).

II.3.1.2 ANÁLISE SINTÁTICA

Para reconhecer um objeto é necessário verificar se a organização de seus vários componentes está correta. Esta verificação é chamada de análise sintática, pois em Grego significa montagem. A análise sintática pode ser definida como o estudo minucioso da montagem e do modo de colocar as várias partes de um todo em posições adequadas (ARARIBÓIA, 1988).

Segundo Câmara Júnior (2007), a análise sintática é aquela que se responsabiliza pela organização do conjunto das palavras, ou seja, as orações. A sintaxe é o conjunto das regras por meio das quais é possível reconhecer a estrutura da frase e as funções de cada um de seus constituintes. Esse procedimento é realizado por meio de um *parser* e o resultado é armazenado em uma árvore.

O analisador sintático utiliza a gramática da Linguagem Natural e cruza com as informações do analisador morfológico, construindo uma árvore de derivação (Figura 2) para a sentença analisada, mostrando o relacionamento das palavras entre si. A árvore

de derivação tem como função converter a lista de palavras que formam a frase em uma estrutura hierárquica representativa de cada palavra da frase separadamente (OLIVEIRA, 2004).

De acordo com Specia (2000, citado por VINHAES, 2005, p. 24) as unidades que foram geradas, correspondem aos componentes que serão atribuídos aos significados quando realizada a análise semântica. Esta contribui para diminuir a quantidade de componentes que a Análise Semântica processará, reduzindo a complexidade do sistema, considerando que o processamento sintático utiliza menos recurso computacional do que o processamento semântico.

Percebe-se na Figura 2 que a frase “*Eu quero imprimir o arquivo .init do Mário*” foi convertida em uma estrutura hierárquica, contendo unidades significativas imprescindíveis à análise semântica (OLIVEIRA, 2004).

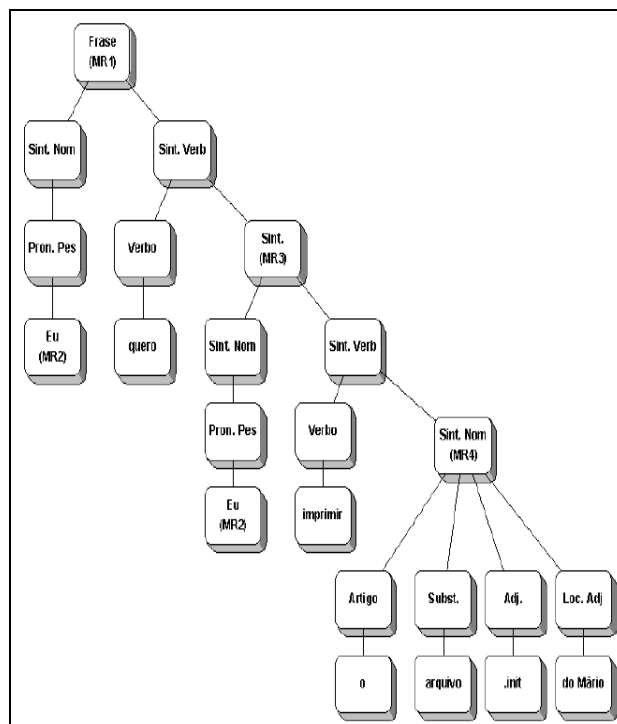


Figura 2 – Resultado da Análise Sintática.
Fonte: OLIVEIRA, 2004.

II.3.1.3 ANÁLISE SEMÂNTICA

A expressão análise semântica tem origem grega e significa dissecação (estudo minucioso) de significado. Tem como objetivo descobrir estratégia para representar o sentido das frases ou sentenças (VINHAES, 2005).

A análise semântica procura analisar as frases sintaticamente corretas para avaliar se as mesmas são compreensíveis. O objetivo é formalizar a interpretação do texto. Não é possível dar significado ao conteúdo, mas é possível analisar as relações válidas entre as palavras, a partir de seus conceitos (CÂMARA JÚNIOR, 2007).

O analisador semântico tem como função verificar o sentido da estrutura das palavras que foram reagrupadas pelo analisador sintático, junto à árvore de derivação (árvore sintática), construída com as informações do analisador morfológico e sintático. (VINHAES, 2005).

Um sistema com capacidade de interpretação semântica avançada conseguiria solucionar a maioria dos casos de ambiguidade encontrados nestas sentenças, visto que, não há regra sintática que solucionaria estes problemas. Um sistema com tal capacidade teria um grande problema, que é o alto grau de processamento de informações, para que pudesse obter a melhor aceção na tradução. Para isso, a consulta em outras bases de dados seria indispensável, com a finalidade de encontrar a resposta satisfatória (OLIVEIRA, 2004).

II.3.1.4 TRADUÇÃO SQL

A etiquetagem da sentença em português consiste na divisão das frases em *tokens*, para que esses possam ser classificados de acordo com as classes gramaticais da língua portuguesa.

Após a etiquetagem da sentença em português, esta etapa tem como objetivo realizar a tradução da frase para o SQL, utilizando o documento XML que contém as informações

(sinônimos) dos nomes das tabelas e dos campos.

II.4 XML

XML (Extensible Markup Language) é a linguagem de marcação de dados (meta-markup language) que provê um formato para descrever dados estruturados. Isso facilita declarações mais precisas do conteúdo e resultados mais significativos de busca através de múltiplas plataformas. O XML também vai permitir o surgimento de uma nova geração de aplicações de manipulação e visualização de dados via internet (DUARTE, 2009).

Os documentos XML possuem uma estrutura básica, composta de um cabeçalho, chamado de prólogo, e do restante do documento, chamado de instância. No prólogo há informações que identificam o documento como sendo XML. A instância do documento segue o prólogo, contendo os dados propriamente ditos, organizados como uma hierarquia de elementos. Os elementos de um documento XML são representados por marcas (*tags*), uma inicial e uma final, podendo possuir conteúdo ou não entre estas marcações. Os elementos podem também possuir atributos, que descrevem suas propriedades e ficam contidos na marca inicial do elemento (GARCIA; MENEZES; BOSCARIOLI, 2008).

A tecnologia XML assemelha-se à tecnologia de Sistemas Gerenciadores de Banco de Dados (SGBDs), uma vez que possui um formato de armazenamento específico (documento) e linguagens de consulta, permite a definição de esquemas e define interfaces (APIs) para o acesso a dados, dentre outras características. Entretanto, a tecnologia XML não é equivalente à tecnologia de SGBD, pois não existem soluções consolidadas para todos os aspectos de gerenciamento de dados, como controle de integridade, e gerenciamento eficiente de transações (LIMA; SCHROEDER; MELLO, 2008).



III DESENVOLVIMENTO

O desenvolvimento deste projeto foi realizado em etapas apresentadas nas seções seguintes.

III.1 AMOSTRA

Para se chegar aos padrões das frases interrogativas foram realizadas várias análises com o intuito de definir as regras das possíveis consultas dos usuários aos dados em linguagem natural. Foram usados como base de teste, um banco de dados de Sistema de Controle de Biblioteca (FIG.3) e um banco de dados de Sistema de Gestão de Centros de Custos (FIG.4).

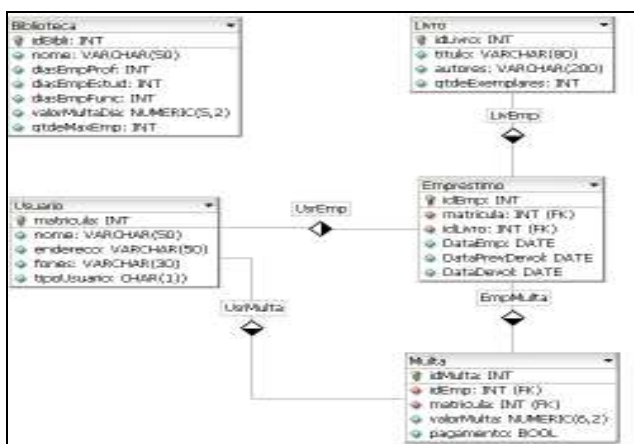


Figura 3 – Esquema de Dados – Sistema de Controle de Biblioteca.

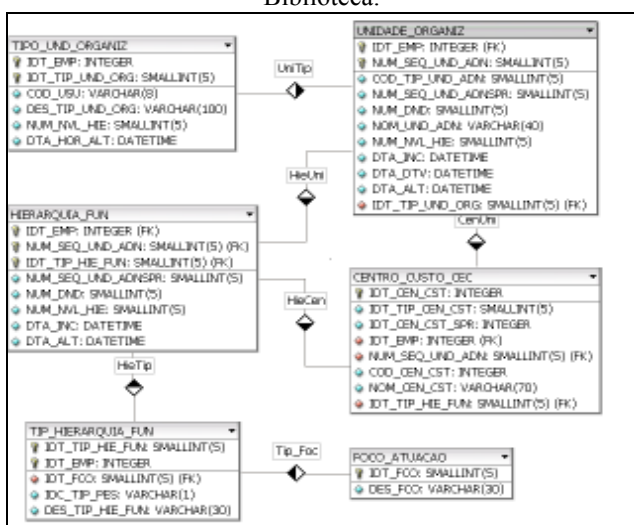


Figura 4 – Esquema de Dados – Sistema de Gestão de Centros de Custos.

III.2 DEFINIÇÃO DO DOMÍNIO

Esta etapa se fez necessário com o intuito de permitir a utilização desta ferramenta com bases de domínios diversificados, onde o usuário escolhe o esquema da base de dados e define sua consulta. Esta escolha consiste nos esquemas dos bancos de dados SQL SERVER 2005 configurado nesta estação de trabalho.

III.3 CONFIGURAÇÃO DO DOMÍNIO

Para que seja possível a identificação dos reais nomes dos atributos e das tabelas, cujos usuários desejam referir, ao discorrerem a consulta, se fez necessária a utilização de um arquivo XML, contendo os sinônimos destes, com o objetivo de isentar o usuário de ter que conhecer particularidades de um modelo de dados no momento da consulta. Desta forma, de acordo com o modelo de dados escolhido são apresentados ao usuário todas as tabelas e atributos ali existentes, para que ele possa definir todos os sinônimos.

Este XML é utilizado no processo de conversão da sentença em português para o SQL.

III.4 ANÁLISE DA SENTENÇA EM LINGUAGEM NATURAL

Esta etapa consiste na análise da sentença onde todas as palavras (*tokens*) são reconhecidas pelo domínio da língua fonte. É nesta etapa que são realizadas as análises morfológicas, sintáticas e semânticas.

Para realizar a etiquetagem da sentença foi utilizado o Módulo Etiquetador desenvolvido em (NUNES, 2007). Foram realizadas algumas modificações neste modulo visto que o mesmo foi desenvolvido para a plataforma web.

No processo de etiquetagem as frases são divididas em *tokens* para que estes possam ser classificados de acordo com as classes

gramaticais da língua portuguesa. A Figura 5 mostra a frase etiquetada (NANTES, 2008).

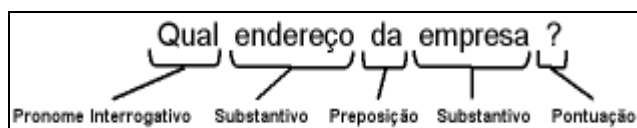


Figura 5 – Exemplo de frase etiquetada.
Fonte: (NANTES, 2008).

Para efetuar o processo de etiquetagem das palavras é necessário possuir informações a respeito das classes gramaticais das palavras a serem etiquetadas. Além disso, é necessário descobrir qual a classe gramatical de uma determinada palavra em um dado contexto, quando essa palavra pertencer a mais de uma classe gramatical (NUNES, 2007).

Para que o etiquetador funcione é necessário que todas as palavras da frase a serem etiquetadas existam na base de dados. Na inexistência da palavra, bem como sua classificação, será concedido ao usuário a opção de incluí-la no dicionário de dados.

Foi modelada e desenvolvida uma base de dados *Dicionário*, que armazena todas as informações necessárias para a etiquetagem das palavras. O Dicionário possui entrada para as unidades léxicas e suas respectivas classificações, assim como as probabilidades para inferências de ambiguidade e não ocorrência.

Esta base de dados Dicionário é utilizada em paralelo com a base do domínio escolhido pelo usuário, permitindo assim a flexibilização da utilização desta ferramenta com outros domínios, tornando independentes as bases de dados. A Arquitetura da solução está definida na sessão 4, onde consta a arquitetura do projeto.

A modelagem desta base de dados em separado se fez necessária com o intuito de isentar o usuário de cadastrar a mesma palavra em domínios diferentes.

Neste módulo foi desenvolvido um etiquetador probabilístico baseado nos conceitos

de QTag de Oliver Mason (2007), projetado para etiquetar textos em inglês (NUNES, 2007).

Quando a palavra pertencer a mais de uma classe gramatical, a ferramenta invoca este método probabilístico. Como exemplo, a unidade léxica “furto”: se a “furto” vier precedida de um pronome e sucedida por um substantivo, por exemplo, como em “eu furto veículos”, a ferramenta avalia a probabilidade de se tratar de um verbo, e a classifica como tal; se a mesma unidade estiver entre um artigo e um adjetivo, como em “o furto planejado”, então será classificada como substantivo. Com esse tipo de análise, as ambiguidades são resolvidas e unidades léxicas não conhecidas podem ser classificadas probabilisticamente.

O QTag é um etiquetador morfológico construído genericamente para qualquer língua, com o objetivo de realizar o Processamento de Linguagem Natural dos textos. A ferramenta é um software livre, desenvolvida pela Universidade de Birmingham que está disponível em <http://www.english.bham.ac.uk/staff/omason/software/qtag.html>, e encontra-se no catálogo da biblioteca digital Linguateca, no endereço eletrônico <http://www.linguateca.pt> (CÂMARA JÚNIOR, 2007).

Esta biblioteca digital é um centro de recursos para processamento computacional da língua portuguesa. Muitos projetos estão disponíveis para utilização, bem como fóruns de discussão para a comunidade de pesquisa da área. Dessa forma, distribui-se o conhecimento e fomenta-se a pesquisa de PLN para o português (CÂMARA JÚNIOR, 2007).

A lista de todas as etiquetas pode ser visualizada na Tabela 1. Cada etiqueta representa uma ou mais classes gramaticais do português (NANTES, 2008).

Observa-se que foi criada uma classe para pontuação o que, via de regra, não é uma classe morfológica. Isso é justificado pelo fato

de que como o processo de tokenização é bastante complexo, optou-se por deixar essa classe para etiquetar caracteres especiais que eventualmente passem pelo crivo do tokenizador.

Tabela 1

ETIQUETAS E GRUPOS

ARTD	Artigos Definidos
ARTI	Artigos Indefinidos
NUM	Numerais Ordinais e Cardinais
ADJ	Adjetivos
PART	Verbos no Particípio
VIRG	Vírgula
PONT	Pontuações
CONJ	Conjunções
ADV	Advérbios
PREP	Preposições
VERB	Verbos
SUB	Substantivos
PRPE	Pronomes Pessoais
PRDE	Pronomes Demonstrativos
PRIN	Pronomes Indefinidos
PRPO	Pronomes Possessivos
PRRE	Pronomes Relativos
PRIT	Pronomes Interrogativos
LOC	Locuções

Fonte: (NANTES, 2008).

III.5 GERAÇÃO DA CONSULTA

Após a geração do arquivo XML, contendo todos os sinônimos das tabelas, dos campos e com todas as palavras já etiquetadas, é dado início o processo de conversão da sentença em português para o SQL. Foram desenvolvidas várias regras com o objetivo de identificar os padrões interrogativos das consultas.

III.5.1 REGRAS DE TRADUÇÃO

Foram analisadas várias consultas nas duas bases de teste (Sistema de Controle de Biblioteca e Sistema de Gestão de Centros de Custos), com o intuito de modelar algumas regras, utilizadas no processo de conversão da sentença em português para o SQL.

As regras foram modeladas de forma aleatórias e de acordo com supostas consultas válidas realizadas pelos usuários, que tinha como objetivo a descoberta dos padrões nas frases. Foi observado as sequencias de tokens para associá-los as cláusulas do select, from e where.

Estas regras são válidas para quaisquer bancos de dados, permitindo assim a flexibilidade da utilização desta ferramenta com qualquer outro modelo relacional.

Segue alguns exemplos onde é mostrada a utilização das regras na base do Sistema de Controle de Biblioteca e logo em seguida a mesma regra, porém aplicada na base do Sistema de Gestão de Centros de Custos.

Quais os livros do autor 'Bruno'?

Após a etiquetagem pelo Módulo Etiquetador, a sentença apresenta as seguintes etiquetas:

Quais os livros do autor 'Bruno' ?
PRIT ARTD SUB PREP SUB SUB PONT

Utilizando o XML com os sinônimos, em SQL esta sentença será convertida para:

```
SELECT LIVRO.TITULO
FROM LIVRO
WHERE LIVRO.AUTORES = 'BRUNO'
```

Para qualquer frase que se enquadrar nesta sequência de etiquetas, esta regra será válida.

pt1 at1 sb1 pp1 sb2 sb3 ?
PRIT ARTD SUB PREP SUB SUB PONT

```
SELECT Sc(sb1)
FROM St(sb1)
WHERE Sc(sb2) = 'sb3'
```

Onde Sc(x), retorna o nome qualificado da coluna (atributo) cujo conjunto de sinônimos contenham x, e St(y), retorna nome da tabela cujo conjunto de sinônimos contenham y.

Segue exemplo da aplicação da mesma regra no banco de dados do Sistema de Gestão de Centros de Custos:

Qual o nível da unidade 'Matriz'?

Qual o nível da unidade 'Matriz' ?
PRIT ARTD SUB PREP SUB SUB PONT

```
SELECT UNIDADE_ORGANIZ.IDT_TIP_UND_ORG
FROM UNIDADE_ORGANIZ
WHERE UNIDADE_ORGANIZ.NOM_UND_ADN
='MATRIZ'
```

pt1 at1 sb1 pp1 sb2 sb3 ?
PRIT ARTD SUB PREP SUB SUB PONT

```
SELECT Sc(sb1)
FROM St(sb2)
WHERE Sc(sb2) = 'sb3'
```

Qual o valor da multa do empréstimo 15?

Qual o valor da multa do empréstimo 15 ?
PRIT ARTD SUB PREP SUB PREP SUB NUM PONT

```
SELECT MULTA.VALORMULTA
FROM MULTA, EMPRESTIMO
WHERE EMPRESTIMO.IDEMP = 15
AND EMPRESTIMO.MATRICULA =
MULTA.MATRICULA
```

pt1 at1 sb1 pp1 sb2 pp2 sb3 nu3 ?
PRIT ARTD SUB PREP SUB PREP SUB NUM PONT

```
SELECT Sc(sb2. sb1)
FROM St(sb2), St(sb3)
WHERE Sc(sb3) = n
AND Sr(sb2 = sb3)
```

Onde Sc(x.y), retorna o nome qualificado da tabela cujo conjunto de sinônimos contenha x e o nome qualificado da coluna (atributo) cujo conjunto sinônimos contenha y. St(w) retorna nome da tabela cujo conjunto de sinônimos contenha w. Sr(w,z) retorna o nome das duas colunas das tabelas w e z que se relacionam.

Aplicação da mesma regra no banco de dados do Sistema de Gestão de Centros de Custos:

Qual o centro de custo da unidade 7054 ?

Qual o centro de custo da unidade 7054 ?
PRIT ARTD SUB PREP SUB PREP SUB NUM PONT

```
SELECT CENTRO_CUSTO_CEC.COD_CEN_CST
FROM UNIDADE_ORGANIZ, CENTRO_CUSTO_CEC
WHERE
CENTRO_CUSTO_CEC.NUM_SEQ_UND_ADN = 7621
AND UNIDADE_ORGANIZ.NUM_SEQ_UND_ADN =
CENTRO_CUSTO_CEC.NUM_SEQ_UND_ADN
```

pt1 at1 sb1 pp1 sb2 pp2 sb3 nu3 ?
PRIT ARTD SUB PREP SUB PREP SUB NUM PONT

```
SELECT Sc(sb2. sb1)
FROM St(sb2), St(sb3)
WHERE Sc(sb3) = n
AND Sr(sb2 = sb3)
```

Qual livro emprestado possui multa ?

Qual o livro emprestado possui multa ?
PRIT ARTD SUB VERB VERB SUB PONT

```
SELECT LIVRO.TITULO
FROM LIVRO LIVRO, EMPRESTIMO EMPRESTIMO,
MULTA MULTA
WHERE LIVRO.IDLIVRO = EMPRESTIMO.IDLIVRO
AND EMPRESTIMO.IDEMP= MULTA.IDEMP
```

pt1 at1 sb1 vb1 vb2 sb3 ?
PRIT ARTD SUB VERB VERB SUB PONT

```
SELECT Sc(sb1)
FROM St(sb1), St(vb1), St(sb3)
WHERE Sr(sb1 = vb1)
AND Sr(vb1 = sb3)
```

O algoritmo que gera o SQL da consulta toma como base um conjunto de etiquetas. Com a frase etiquetada, o módulo retorna dois vetores: um com a frase dividida pelas palavras e outro com as etiquetas das palavras. O algoritmo percorre o vetor de etiquetas para descobrir em qual das regras que se encaixa. Descoberta a regra, são feitas consultas para encontrar quais tabelas ou campos serão utilizados no SQL. Para isso utilizam-se os sinônimos que estão presentes no XML.

IV IMPLEMENTAÇÃO

Este protótipo de ferramenta foi implementado em linguagem de programação Java, utilizando o NetBeans como ambiente de desenvolvimento. O banco de dados utilizado foi o SQL SERVER 2005.

Nesta etapa é apresentada a arquitetura da solução, bem como o detalhamento das funcionalidades deste projeto.

IV.1 ARQUITETURA DA SOLUÇÃO

A Figura 6 é uma representação dos fluxos dos processos e atividades permitindo um melhor entendimento deste projeto.

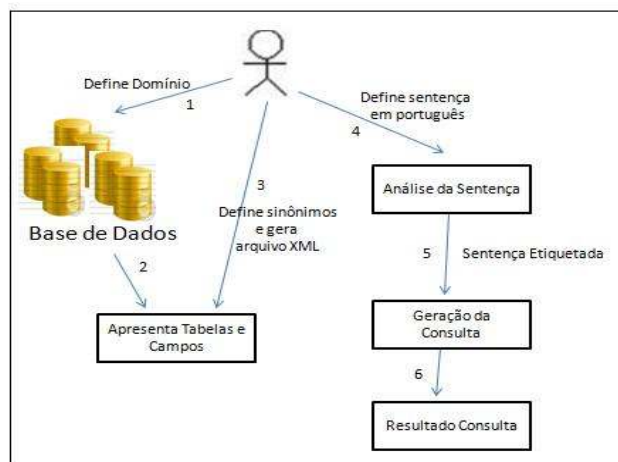


Figura 6- Fluxo dos processos e atividades.

O início do processo tem como autor o usuário que define o domínio (passo 1) onde deseja discorrer com sua sentença em português

por meio da escolha de uma base de dados disponível. Feita esta escolha, é apresentado ao usuário todas as tabelas e campos para que o mesmo defina os sinônimos e gere o arquivo XML (passo 2).

O usuário define a sentença em português (passo 4) passando a frase para o módulo de análise da sentença para que este possa etiquetar a frase. A sequência de etiquetas é enviada ao módulo de geração de consultas (passo 5), que tem como objetivo converter a sentença em SQL e retornar o devido resultado da consulta (passo 6).

Este projeto foi desenvolvido seguindo o padrão *Model-View-Controller* (MVC), que é um padrão de arquitetura de software, cuja premissa é isolar a aplicação em camadas distintas. Os componentes deste projeto foram divididos em três camadas: Apresentação, Camada de Negócios e Dados.

- **Apresentação:** É a camada de interface com o usuário, nela foram criadas três interfaces que permite ao usuário definir a base de dados (domínio), a consulta, visualizar a sentença etiquetada e o resultado da consulta. Com esta camada, quaisquer alterações feitas no layout não afetam a manipulação de dados.
- **Camada de Negócios:** Camada que trata o negócio em si, definindo o comportamento da aplicação. Contém todo o processamento da Linguagem Natural.
- **Dados:** Camada responsável por todos os acessos e tratamentos de dados. É por meio desta que são feitos os acessos aos bancos de dados e ao XML.

Na Figura 7 é ilustrado o Esquema de Dados do Dicionário.

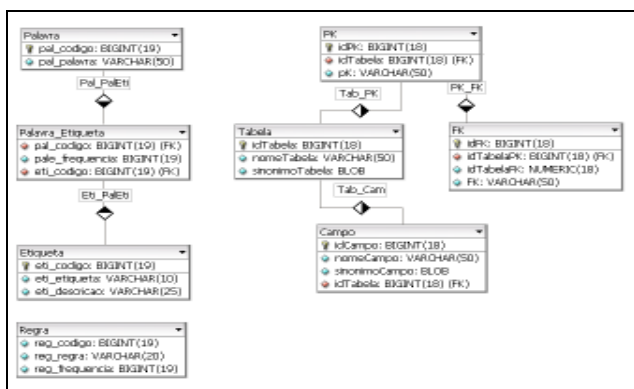


Figura 7- Esquema de Dados da Base de Dados do Dicionário.

O processo utiliza as seguintes tabelas:

- Regra – para inserir regras para se conseguir a etiquetagem das palavras e as ocorrências dessas regras;
- Palavra – onde são inseridas todas as palavras que tiveram ocorrência no sistema;
- Etiqueta – onde foram armazenadas todas as etiquetas definidas para o sistema;
- Palavra_etiqueta – utilizada para armazenar a palavra com sua etiqueta e quantas ocorrências esta palavra teve no sistema;
- Tabela – onde são armazenadas todas as tabelas de um determinado banco de dados, que é a base das informações que se deseja obter e os sinônimos destas.
- Campo – armazena todos os campos de uma determinada tabela e seus sinônimos.
- PK – que contém as chaves primárias de todas as tabelas.
- FK – onde está armazenada a relação das chaves estrangeiras que se relacionam com outras chaves primárias, bem como os atributos que se relacionam.

IV.2 DETALHES DE IMPLEMENTAÇÃO

Esta seção tem como objetivo a descrição da implementação que será descrita a seguir.

Para que o usuário possa operar o sistema, deve-se escolher o domínio do banco de dados onde constam as informações a serem consultadas conforme mostra a Figura 8.

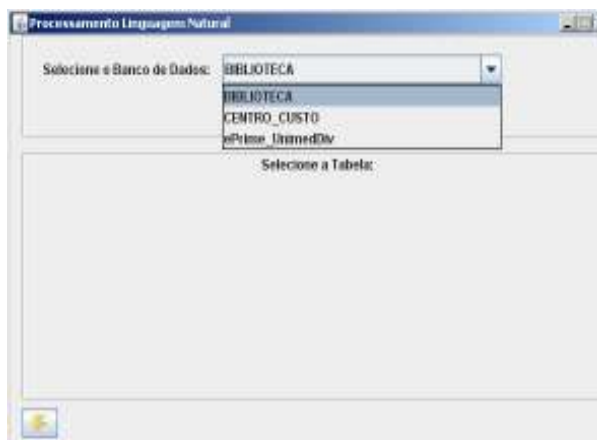


Figura 8- Tela para escolha do banco de dados.

Na sequência são apresentadas ao usuário todas as tabelas para que o mesmo possa configurar o domínio, definindo assim os sinônimos destas (Figura 9).



Figura 9 - Tela para escolha da tabela para definição dos sinônimos.

Após efetuado um clique sobre o nome da tabela desejada, serão listados todos os campos para que se possa definir os sinônimos, tanto da tabela quanto do campo.

Terminada a escrita dos sinônimos deve-se efetuar um clique sobre o botão Gerar XML (Figura 10) para que o sistema gere o arquivo XML com as tabelas, campos e seus sinônimos (Figura 11).



Figura 10 - Tela para definição dos sinônimos da tabela e campos.

```

<Tabela>
  <nome>Livro</nome>
  <Sinônimos>
    <string>livro;livros;autor;autora</string>
  </Sinônimos>
  <Campos>
    <Campo>
      <NomeCampo>Autores</NomeCampo>
      <sinonimoCampos>
        <string>autores;autor;escritor</string>
      </sinonimoCampos>
    </Campo>
    <Campo>
      <NomeCampo>IdLivro</NomeCampo>
      <sinonimoCampos>
        <string>codigo;identificador;numero;codigo do livro</string>
      </sinonimoCampos>
    </Campo>
    <Campo>
      <NomeCampo>QtdeExemplares</NomeCampo>
      <sinonimoCampos>
        <string>quantidade;exemplares;quantidade disponivel;quantidade de exemplares</string>
      </sinonimoCampos>
    </Campo>
    <Campo>
      <NomeCampo>Titulo</NomeCampo>
      <sinonimoCampos>
        <string>titulo;descricao;livro;livros;nome</string>
      </sinonimoCampos>
    </Campo>
  </Campos>
</Tabela>

```

Figura 11 – Arquivo do XML gerado pelo sistema contendo os sinônimos da tabela Livro e seus atributos.

Definido o domínio e os sinônimos, para que o sistema efetue a análise da sentença em Linguagem Natural, o usuário deve digitar uma

sentença em português sobre o que ele deseja consultar em Linguagem Natural e efetuar um clique sobre o botão consultar (Figura 12),

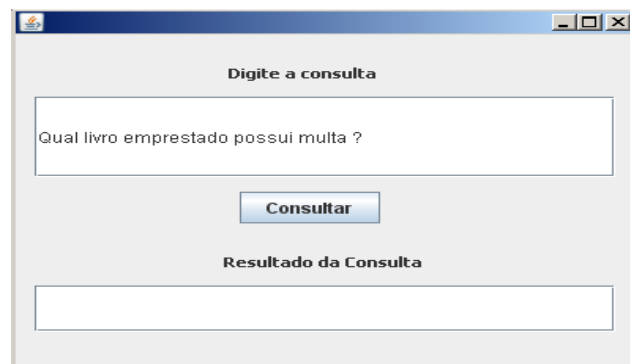


Figura 12 – Tela para consulta em Linguagem Natural.

Depois de inserida a consulta, é feita a etiquetagem da frase. Caso o sistema não reconheça alguma palavra da sentença da consulta, será exibida ao usuário a lista de todas as palavras desta, para que seja definida ou corrigida a etiqueta de cada palavra. Desta forma o sistema poderá treinar o etiquetador do sistema (Figura 13).



Figura 13 – Tela da etiquetagem da frase.

Após a etiquetagem da frase, o usuário deverá efetuar um clique sobre o botão “Continuar...” para que o sistema utilize as regras de tradução e realize uma consulta no banco de dados, retornando assim uma resposta precisa sobre o que foi pesquisado (Figura 14).

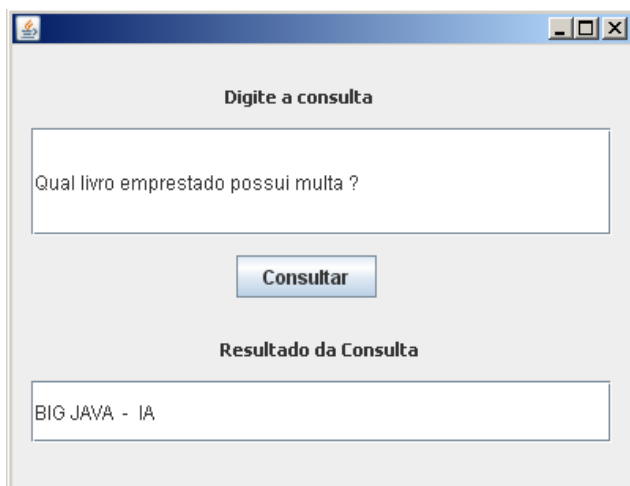


Figura 14 – Tela de consulta com a resposta.

V ANÁLISE DE RESULTADOS

O resultado obtido neste projeto foi a geração de um documento XML estruturado para obter informações como nomes e sinônimos das tabelas e campos. A partir deste XML e da sentença etiquetada é possível a geração da instrução SQL, bem como retornar o resultado da mesma.

Para análise dos resultados foram inseridas três consultas de acordo com o Modelo Entidade Relacionamento – Sistema de Controle de Biblioteca (Figura 3) que obtiveram os resultados esperados e uma que não atendeu, visto que a frase não estava semanticamente correta.

Entrada da sentença:

Qual o nome do usuário da matricula 405218109 ?

Sentença etiquetada:

PRIT ARTD SUB PREP SUB PREP SUB NUM PONT

Consulta em SQL:

```
SELECT USUARIO.NOME FROM USUARIO WHERE
USUARIO.MATRICULA = 405218109
```

Resultado da sentença:

BRUNO PERCHÉ PINTO

Entrada da sentença:

Qual o nome do livro do emprestimo para matricula 405218109 ?

Sentença etiquetada:

PRIT ARTD SUB PREP SUB PREP SUB PREP SUB NUM PONT

Consulta em SQL:

```
SELECT LIVRO.TITULO FROM LIVRO LIVRO,
EMPRESTIMO EMPRESTIMO WHERE
EMPRESTIMO.MATRICULA = 405218109 AND
LIVRO.IDLIVRO = EMPRESTIMO.IDLIVRO
```

Resultado da sentença:

BIG JAVA

Entrada da sentença:

Qual o valor da multa da usuaria 'RAQUEL' ?

Sentença etiquetada:

PRIT ARTD SUB PREP SUB PREP SUB SUB PONT

Consulta em SQL:

```
SELECT MULTA.VALORMULTA FROM USUARIO
USUARIO ,MULTA MULTA WHERE
USUARIO.MATRICULA= MULTA.MATRICULA
```

Resultado da sentença:

2,00

Entrada da sentença:

Qual o nome do prof da matricula qual?

Sentença etiquetada:

PRIT ARTD SUB PREP SUB PREP SUB PRIT PONT

Consulta em SQL:

“”

Resultado da sentença:

EXISTE UM ERRO SEMANTICO NA FRASE

Os testes realizados atenderam as expectativas, retornando as devidas informações. O tratamento da Linguagem Natural no Português é muito complexo, visto que a língua portuguesa é livre de contexto, sendo composta por várias regras e por várias exceções, dificultando assim a análise da sentença.

Normalmente computadores estão aptos a compreender instruções escritas em linguagens de computação como o Java, C, PERL, Basic, etc., mas possuem muita dificuldade em entender comandos escritos em uma linguagem humana. Isso se deve ao fato das linguagens de computação serem extremamente precisas,

contendo regras fixas e estruturas lógicas bem definidas que permitem o computador saber exatamente como deve proceder a cada comando. Já em um idioma humano uma simples frase normalmente contém ambiguidades e interpretações que dependem do contexto, do conhecimento do mundo, de regras gramaticais, culturais e de conceitos abstratos.

VI TRABALHOS FUTUROS

Como trabalhos futuros podem ser tratadas consultas mais complexas e tornar o sistema o mais genérico possível. Também poderá ser incorporado um dicionário da língua portuguesa onde as palavras poderiam ser armazenadas em sua forma canônica, mais básica, e que o sistema seja capaz de derivar as formas mais complexas de cada entrada. Poderão ser realizados testes com modelos ER mais complexos.

Considera-se que a complexidade do problema deve proporcionar novos estudos que venham a contribuir para o aprimoramento do processo de decisão relativo à seleção de termos que vão compor as estratégias de busca em bases de dados.

VII CONCLUSÃO

Este projeto mostrou que o Processamento de Linguagem Natural pode trazer benefícios para implementações de aplicações inteligentes, com maior interação do usuário com a máquina o mais natural possível, onde usuários podem extrair informações sem a necessidade de conhecer particularidades de um modelo de dados ou de estruturas de armazenamento.

A flexibilização das consultas é a principal vantagem da utilização da Linguagem Natural, livrando o usuário, principalmente os leigos, de ter que dominar linguagens formais de consulta como o SQL, ou de ter que estar limitado a rígidas opções de consultas pré-determinadas.

Não se sabe se um dia os computadores poderão igualar (ou superar) a capacidade humana de entender ou compor textos. Atualmente estas capacidades são bastante limitadas no computador, mas muitos resultados práticos já são possíveis e utilizados por diversos tipos de programas.

O estudo de PLN em línguas com morfologia e sintaxe reconhecidamente difíceis, como português ou francês, é um desafio que se impõe para pesquisadores da área de Inteligência Artificial. Entretanto, inúmeras aplicações terão benefícios com a evolução desses resultados. A recuperação de informação em banco de dados por meio da Linguagem Natural, bem como diversas outras áreas de pesquisa em inteligência artificial, não tem o objetivo de substituir o trabalho humano, mas sim garantir níveis mínimos de atendimento a demandas de trabalho não estritamente intelectual.

Permite-se destarte, que os esforços sejam efetivamente engendrados nos problemas que exijam cognição humana, melhorando por fim, a qualidade de vida das pessoas. O controle exercido por máquinas, longe de escravizar os seres humanos, vai libertar homens e mulheres para realização de tarefas que somente eles podem fazer.

AGRADECIMENTOS

A todos aqueles que, de certa forma, contribuíram para a realização deste trabalho, em especial, à Prof.^a Magali Barroso, pela minuciosa revisão do texto.

REFERÊNCIAS

- ARARIBÓIA, G. **Inteligência Artificial**: Um curso prático. Rio de Janeiro: Livros Técnicos e Científicos, 1988. 286 p.
- CÂMARA JÚNIOR, Auto Tavares. **Indexação Automática de Acordãos por Meio de Processamento de Linguagem Natural**. 2007.



- 142f. Dissertação de Mestrado – Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação, Universidade de Brasília. Brasília, 2007.
- CARDOSO, Olinda Nogueira Paes. **Recuperação de Informação**. Lavras, 2007.
- CROFT, William. B. (Editor) **Advances in Information Retrieval**. London: Kluwer Academic Publishers, 2000.
- DUARTE, Otto Carlos Muniz Bandeira. **XML - Extensible Markup Language**. Universidade Federal do Rio de Janeiro, 2009. Disponível em: http://www.gta.ufjf.br/grad/00_1/miguel/index.html. Acesso em: 23 nov. 2009.
- DUQUE, Cláudio Gottschalg. **SiRILiCO Uma Proposta para um Sistema de Recuperação de Informação baseado em Teorias da Linguística Computacional e Ontologia**. 2005. 120 f. Tese de Doutorado em Ciência da Informação – Escola de Ciência da Informação, Universidade Federal de Minas Gerais. Belo Horizonte.
- FERNANDES, A. M. R. **Inteligência Artificial: noções gerais**. Florianópolis: Visual Books tda, 2003.
- GARCIA, Maxiwell Salvador; MENEZES, Anderson Luiz; BOSCARIOLI, Clodis. **Um Estudo sobre Gerenciamento de Dados XML em SGBDs Objeto-Relacionais e em SGBDs com suporte nativo a XML**. Cascavel, 2007.
- JACQUEMIN, Christian. **Traitement Automatique des Langues pour la Recherche d'Information**. Paris: Hermès Science Publications, 2000. 591 p.
- LIMA, Claudio; SCHROEDER, Rebeca; MELLO, Ronaldo Santos. **Uma Ferramenta para Conversão de Esquemas Conceituais EER para Esquemas Lógicos XML**. Florianópolis, 2008.
- MASON, Oliver. **Qtag**. Universidade de Birmingham no Reino Unido, 2007. Disponível em: <http://www.english.bham.ac.uk/staff/omason/software/qtag.html>. Acesso em: 19/04/2010.
- MIORELLI, Sandra Teresinha. **ED-CER: Extração do Sintagma Nominal em Sentenças em Português**. 2001. 98f. Dissertação de Mestrado em Ciência da Computação - Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul.
- MURADAS, Alfredo Martins. **Interface em Linguagem Natural para Consulta a Bancos de Dados Relacionais**. 1995. 162f. Dissertação de Mestrado Em Ciências em Sistemas e Computação - Instituto Militar de Engenharia. Rio de Janeiro, 1995.
- NANTES, Leandro Martins. **Desenvolvimento de um Sistema Baseado em Linguagem Natural para Consultas em Banco de Dados na Web**. Presidente Prudente, 2008.
- NUNES, RODRIGO ALVES. **Sistema Baseado em Linguagem Natural para a Recuperação de Imagens em um Banco de Imagens**. Presidente Prudente, 2007.
- OLIVEIRA, Fábio Abreu Dias. **Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa**. Revista de Ciência da Informação. Rio Grande do Sul, 2004.
- RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**. Rio de Janeiro: Campus, 2004. 331 p.
- SILVA, Renato Rocha; LIMA, Sergio Muinhos Barroso. **Consultas em Bancos de Dados Utilizando Linguagem Natural**. Juiz de Fora, 2007.
- SOUZA, Solange N. A. ; CAMPOS, E. G. L. ; SANTOS, A. R. D. . **Uma ferramenta para Definição de Consultas Baseadas em Papéis e Entidades**. Revista IEEE América Latina, v. 4, p. 277-282, 2006.
- SPECIA, L. Modelagem de um Interpretador Lexical para a Linguagem DART. Cascavel,

2000. apud VINHAES, Reges faria. **Estudo da utilização de processamento de linguagem natural para otimização de tradutores automáticos**. 2005. 57f. Tese de Doutorado em Ciências da Comunicação – Escola de Comunicação e Artes, Universidade de São Paulo.

VARELLA, Amanda *et al.* **Uma interface em Linguagem Natural para a verificação de Regras de Negócios em bases de dados**. Rio de Janeiro, 2006.

VINHAES, Rêges Faria. **Estudo da utilização de processamento de linguagem natural para otimização de tradutores automáticos**. 2005. 57f. Tese de Doutorado em Ciências da Comunicação – Escola de Comunicação e Artes, Universidade de São Paulo.