UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ

CAMPUS LUIZ MENEGHEL

RICARDO BARBOSA CRIVELLI

RECUPERAÇÃO DE INFORMAÇÃO POR MEIO DE PROCESSAMENTO DE LINGUAGEM NATURAL

Bandeirantes

2011

RICARDO BARBOSA CRIVELLI

RECUPERAÇÃO DA INFORMAÇÃO POR MEIO DE PROCESSAMENTO DE LINGUAGEM NATURAL

Monografia apresentada à Universidade Estadual do Norte do Paraná – *campus* Luiz Meneghel – como para obtenção do grau de Bacharel em II de Sistemas de Informação.

Orientador: Prof. Me. Glauco Carlos Silva

Bandeirantes

2011

RICARDO BARBOSA CRIVELLI

RECUPERAÇÃO DA INFORMAÇÃO POR MEIO DO PROCESSAMENTO DE LINGUAGEM NATURAL

Monografia apresentada à Universidade Estadual do Norte do Paraná – *campus* Luiz Meneghel – como requisito para obtenção do grau de Bacharel em Sistemas de Informação.

COMISSÃO EXAMINADORA

_____
Prof. Me. Glauco Carlos Silva
UENP – *Campus* Luiz Meneghel

_____
Prof. Me. André Luis Andrade Menolli
UENP – *Campus* Luiz Meneghel

_____
Prof. Me. José Reinaldo Merlin
UENP - *Campus* Luiz Meneghel

Bandeirantes, 07 de dezembro de 2011

Dedico este trabalho aos meus pais, Pedro e Tereza. Amo vocês.

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Pedro e Tereza e minha irmã, Cristiane por acreditarem e financiarem este sonho.

À minha namorada, Michaella, por me dar forças, aguentar a distância, me amar acima de tudo e ser esse pilar fundamental em minha vida.

Agradeço também ao meu orientador, Glauco por me auxiliar e me nortear nessa importante etapa da minha vida.

Ao meu irmão e companheiro de luta, Thiago (Fei) por todos os momentos que passamos nesses anos, você se tornou mais que um companheiro de república, se tornou um irmão, ao Virgílio (Buuh) que foi além de um companheiro de república, um grande amigo e professor sobre bonsai, à todos da República Kantagalo, pelas cervejadas, brincadeiras e madrugadas acordados, à Tia Vera por ser mais que uma sogra, ser uma segunda mãe, à D. Rosinha por preparar doces maravilhosos e me acolher, ao Núcleo de Tecnologia de Informação da UENP e todos os seus funcionários e estagiários, ao pessoal da universidade e colegas de sala, à toda família de minha namorada (especialmente a minha irmã Gabi) aos meus professores, por compartilharem o seus bens mais preciosos, o conhecimento, e por último, mas não menos importante, aos meus sócios Fabrício (Mano) e ao Diego (Dieguito) pelas horas trabalhadas, pelas reuniões de trabalho no bar, pelas discussões, pela luta e pelo companheirismo.

E a todos que por ventura eu tenha esquecido e a todos que ajudaram de qualquer forma para que esse sonho se tornasse realidade.

Não tentes ser bem
sucedido, tenta antes ser
um homem de valor.
(Albert Einstein)

But lo! Men have
become the tools
of their tools.
(Henry David Thoreau)

RESUMO

Este trabalho apresenta uma breve introdução à técnica de Processamento de Linguagem Natural e suas etapas, análise morfológica, análise sintática, análise semântica e por último a análise pragmática, à técnica de Recuperação da Informação, mostrando o seu funcionamento, como são compostos os Sistemas de Recuperação da Informação e os componentes da arquitetura e funcionamento dos Sistemas de Perguntas e Respostas. Também é apresentada uma proposta de desenvolvimento de um protótipo de sistema de perguntas e respostas automatizado utilizando linguagem natural para busca de horários de ônibus armazenados em um banco de dados. Tal protótipo utiliza a linguagem natural para comunicação homem/máquina à partir da integração e desenvolvimento de ferramentas previamente analisadas e testadas que fazem o processamento de cada etapa do Processamento de Linguagem Natural.

Palavras-chave: Processamento de Linguagem Natural, Sistemas de Perguntas e Respostas.

ABSTRACT

This paper presents a brief introduction to the technique of Natural Language Processing and its stages, morphological analysis, parsing, semantic analysis and finally the pragmatic analysis, the technique of Information Retrieval, showing its operation, how Information Retrieval Systems are composed and the components of the architecture and functioning of Questions and Answers. Also presented is a proposal to develop a automated prototype system of questions and answers using natural language to search for bus schedules stored in a database. This prototype uses a natural language for communication man/machine from the integration and development of previously analyzed and tested tools that make the processing of each step of the Natural Language Processing.

Key words: Natural Language Processing and Questions and Answers Systems.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O Homem utiliza a linguagem para se comunicar desde os seus primórdios e ela é sem dúvida um dos grandes fatores de sucesso para seu desenvolvimento.

Desenvolvimento este, que levou o homem a criar o computador e junto com ele a linguagem de máquina. A maior vantagem da linguagem de máquina é que diferente da linguagem natural não possui ambiguidades e é altamente lógica e formal exigindo que os humanos sejam forçados a aprenderem uma nova linguagem para se comunicarem com o computador, porém com o avanço tecnológico e o aumento do número de pessoas que utilizam o computador a dificuldade de se recuperar a informação contida nos computadores vem se tornando um grave problema e passou-se a estudar ainda mais o processamento de linguagem natural.

> A tarefa de processar uma linguagem natural permite que os seres humanos comuniquem-se com os computadores da forma mais "natural" possível, utilizando a linguagem com a qual mais estão habituados. Elimina-se, desta maneira, a necessidade de adaptação a formas inusitadas de interação, ou mesmo o aprendizado de uma linguagem artificial, cuja sintaxe costuma ser de difícil aprendizado e domínio, a exemplo das linguagens de consulta a bancos de dados. (OLIVEIRA, 2011)

Para que essa comunicação seja feita de uma forma mais simples este trabalho propõe a implementação de um protótipo de um Sistema de Perguntas e Respostas que utiliza o Processamento de Linguagem Natural, por meio de suas etapas, para interpretar o que o usuário perguntou e da Recuperação da Informação buscar em um banco de dados a resposta.

Após esse processo deverá fazer o processo inverso passando pelas etapas do Processamento de Linguagem Natural para informar a resposta para a pergunta para o usuário no idioma Português.

O Processamento de Linguagem Natural é uma área multidisciplinar, envolvendo principalmente a Psicologia Cognitiva, a Computação e a Linguística (ROSA, 1995) e possui cinco etapas, a análise da voz, que não será utilizada nesse trabalho, pois a entrada do usuário será feita via texto, a análise morfológica, a análise sintática, a análise semântica e por último a análise pragmática e tem como objetivo

transformar o texto em linguagem natural, neste caso a língua portuguesa em linguagem de máquina e vice-versa.

Conhecido por RI ou *information retrieval* (IR), a recuperação da informação surgiu em 1950 e tem como meta encontrar a informação exigida para satisfazer a necessidade de informação (NI) do usuário (FRANTZ 1997 apud GONZALEZ et al.) através da busca em uma ontologia, normalmente armazenada em um banco de dados.

1.1 CONTEXTO E DELIMITAÇÃO DO TRABALHO

Um protótipo de sistema de interpretação de perguntas para um domínio específico que retornará as respostas embasadas em um banco de dados previamente construído.

1.2 FORMULAÇÃO DO PROBLEMA

Com a tecnologia disponível atualmente, não é necessário que o homem aprenda a linguagem de máquina que é altamente lógica e complexa para se comunicar com o computador, ele deve ser capaz de se comunicar da mesma maneira que se comunica com as outras pessoas, por meio da linguagem natural. Para tanto a proposta deste trabalho é um protótipo de ferramenta de perguntas e respostas onde a entrada enviada pelo homem ao computador e a resposta a emitida pela máquina sejam em sua linguagem natural possa melhorar a interação entre as partes.

1.3 OBJETIVOS

O objetivo deste trabalho é construir um protótipo de sistema processamento de linguagem natural que seja capaz de compreender perguntas enviadas em linguagem natural pelo usuário com o uso do processamento de linguagem natural e respondê-las buscando em um dicionário também na língua em sua forma natural.

1.3.1 Objetivos Específicos

- Estudo sobre Processamento de Linguagem Natural
- Estudo sobre Recuperação da Informação
- Estudo sobre sistemas de Perguntas e Respostas
- Pesquisar e comparar as ferramentas já existentes;
- Adaptar, Integrar e/ou desenvolver um protótipo de ferramenta para o processo;
- Aplicar o protótipo a uma base de conhecimento para recuperação da informação
- Testar a solução proposta.

1.4 JUSTIFICATIVA

Nas últimas décadas tem-se observado grande aumento na quantidade de informação armazenada e disponibilizada em documentos, principalmente eletrônicos (WIVES, 2004), tornando o acesso às informações que são realmente importantes mais difícil.

> Perante um enorme volume de dados é imprescindível um processo de os organizar para se conseguir o melhor proveito. São necessárias técnicas automáticas de filtragem de informação, como Sistemas de Recuperação de Informação (SRI), para auxiliar o utilizador a encontrar aquilo que realmente procura (SAIAS, 2003).

Um outro problema encontrado é como essas informações serão acessadas, pois elas exigem o conhecimento das linguagens estruturadas, comumente chamadas de linguagens de programação e muitas são as dificuldades encontradas para o aprendizado desta linguagem. Segundo MENDES (2011), a maior dificuldade do aprendizado de uma linguagem de programação diz respeito à concepção e formalização de uma solução para um determinado problema e não à sua codificação.

1.5 METODOLOGIA

Para a obtenção do conhecimento sobre o Processamento de Linguagem Natural, Recuperação da Informação e sobre os Sistemas de Perguntas e Respostas foi feito um estudo em artigos e monografias, além de livros e periódicos sobre o assunto.

As ferramentas necessárias para construção do protótipo foram encontradas à partir de buscas na internet e em trabalhos publicados e posteriormente foram submetidas à testes iguais e seus resultados comparados para a escolha de qual será utilizada, feita essa escolha elas serão integradas e adaptadas para o Português, se necessário para criação de um protótipo de sistema de perguntas e respostas automatizado para busca horários de ônibus que utiliza a linguagem natural para comunicação homem/máquina.

Criado o protótipo, uma base de conhecimento, armazenada em um sistema de gerenciamento de banco de dados, será aplicada para que o protótipo consiga fazer a recuperação das informações requisitadas pelo usuário.

Com o sistema funcionando será feito testes para verificar quantas perguntas ele responde de maneira satisfatória e quais são os principais motivos dos erros.

1.6 ORGANIZAÇÃO DO TRABALHO

O trabalho está organizado da seguinte forma. A Seção 2 apresenta a fundamentação teórica para que o entendimento sobre o desenvolvimento possa ser claro. A Seção 3 apresenta o Processamento de Linguagem Natural, técnica utilizada neste trabalho. A Seção 4 mostra a pesquisa realizada e seus resultados. A Seção 5 apresenta as conclusões da pesquisa e a Seção 6 apresenta os trabalhos futuros para continuidade deste.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo será abordado as principais características do processamento de linguagem natural e suas estruturas bem como uma seção sobre Recuperação da Informação.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

Conhecida também por PLN (ou NLP, em inglês), o Processamento de Linguagens Naturais é um assunto ligado à Ciência da Computação e a outras áreas do conhecimento como linguística, filosofia e psicologia, a qual lida com a capacidade computacional da linguagem humana, a Figura 1 ilustra as principais áreas em que o PLN está ligado.

Por conseguinte, tal área tem por objetivo sistematizar e estabelecer uma metodologia para analisar textos e gerar frases do idioma humano, de forma que possamos nos comunicar com as máquinas de maneira intuitiva e descomplicada como se estivéssemos nos comunicando com outra pessoa. O fato de uma máquina entender uma linguagem natural implica na mesma fazer análises sintáticas, morfológicas, semânticas e léxicas, obter informações, gerar resumos, interpretar o sentido e aprender conceitos novos através de sentenças da linguagem humana.

Atualmente para interagir com os computadores deve-se seguir padrões mais rígidos e não intuitivos de comunicação criando-se, assim, a necessidade do aprendizado de uma linguagem artificial (comumente conhecida como linguagem de máquina), geralmente de assimilação e domínio mais complexo.

2.1.1 Histórico

Os primeiros estudos na área de Processamento de Linguagem Natural tiveram início nos anos 50. Os primeiros sistemas criados nos anos 60 já conseguiam responder de forma rudimentar perguntas enviadas pelo usuário sobre muitos assuntos, como por exemplo, matemática e inglês.

Figura 1 - Áreas às quais o PLN está ligado. Baseado em: SCHNEIDER (2001)

Segundo OLIVEIRA (2011), existem quatro categorias históricas de programas, sendo elas:

- Programas que geram um número reduzido de resultados em domínios específicos, alguns exemplos deste tipo de programa são, o BASEBALL, SAD-SAM e o ELIZA;
- Programas que indexam textos para auxiliar na recuperação de determinadas palavras ou frases. Um ponto fraco destes sistemas é que eles são semanticamente fracos e não tem poderes dedutivos. Um exemplo de sistema indexador é o PROTO-SYNTHEX;
- Programas classificados como de lógica limitada tem por objetivo traduzir as entradas para notação formal usada na base de dados. Estes programas conseguem fazer deduções a partir da informação mantida no banco de dados. Alguns sistemas nesta categoria são o TLC, DEACON e CONVERSE.
- Os programas com base de conhecimento usam as informações sobre um domínio específico para compreender as entradas do usuário. Estes sistemas constituíam sistemas especialistas com grande poder dedutivo.

2.1.2 Etapas do Processamento de Linguagem Natural

O processo de Processamento de Linguagem Natural, como ilustrado na figura 2, pode ser divido em cinco etapas.

2.1.2.1 Análise da Voz

Segundo SCHNEIDER (2001) os sistemas atuais para reconhecimento da voz se dividem em Reconhecimento de Voz Discreta e Reconhecimento de Voz Contínua.

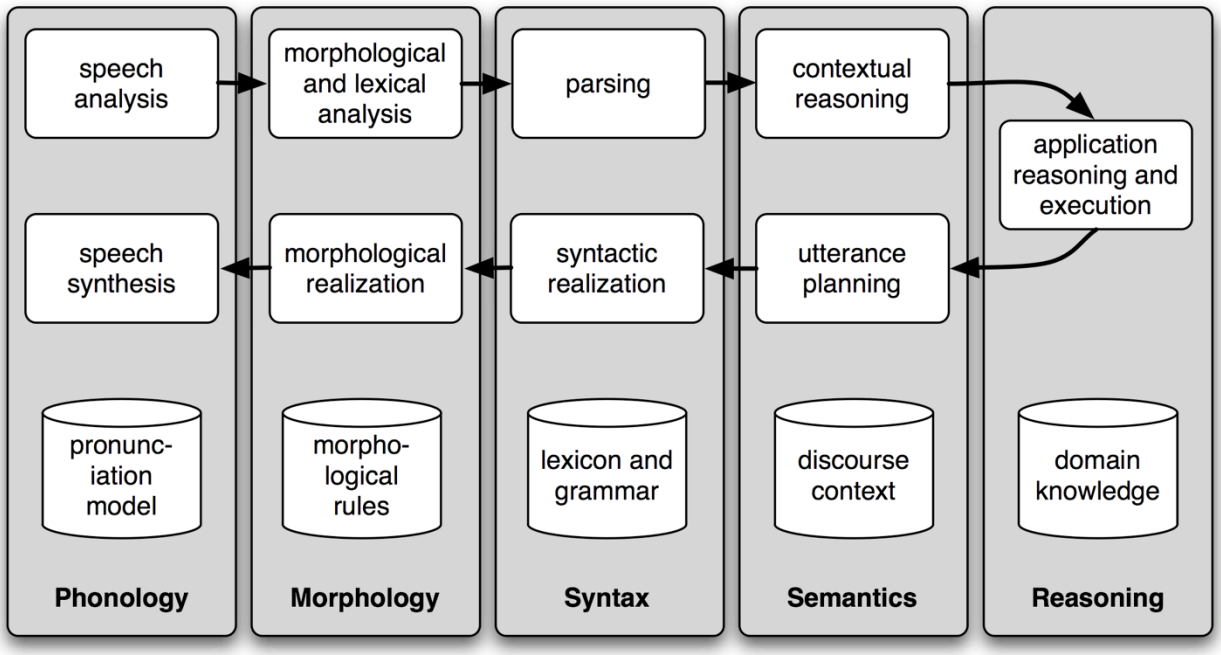

Figura 2 - Etapas do PLN. Fonte: Natural Language Processing with Phyton, p. 32.

2.1.2.1.1 Reconhecimento de Voz Discreta

Os sistemas de reconhecimento de voz discreta são sistemas que requerem que o usuário fale cada palavra separada. Isto traz uma necessidade muito menor de cálculos, porém, é completamente impraticável para sistemas de ditado. Se utiliza estes programa principalmente para fornecer comandos distintos para um computador como em Telebanking, mas também para por exemplo o controle de jogos (LINGUATEC, 2001 apud SCHNEIDER, 2001).

2.1.2.1.2 Reconhecimento de Voz Contínua

Utilizados como uma forma de secretária automática os sistemas de Voz Contínua têm tarefas muito mais complicadas para resolver, pois a separação das palavras em uma frase contínua requer bem mais recursos e soluções tecnológicas inteligentes (LINGUATEC, 2001 apud SCHNEIDER, 2001).

2.1.2.2 Análise Morfológica

Nesta etapa do processamento é feita a separação do prefixo e sufixo das palavras, armazenando-se assim somente seu radical. Um exemplo de prefixo seria o *des* na palavra desnecessário e um sufixo seria o *mente* de tranquilamente.

O tratamento computacional deste tipo de análise é relativamente simples. Baseia-se em regras que analisam as palavras e as classificam segundo tabelas de afixos (MÜLLER, 2003).

Um reconhecedor utilizado para a análise morfológica é o autômato finito. Foi proposta uma forma de comprimir vocabulários extensos através de autômatos determinísticos acíclicos minimizados (OLIVEIRA, 2011).

O emprego do analisador morfológico é fundamental para a compreensão de uma frase, pois para formar uma estrutura coerente de uma sentença, é necessário compreender o significado de cada uma das palavras componentes (RICH, 1993 apud OLIVEIRA, 2011).

2.1.2.3 Análise Sintática

Nesta etapa o analisador sintático cria uma árvore de derivação para cada sentença, mostrando como as palavras estão ligadas entre si.

De acordo com OLIVEIRA, é durante a construção da árvore de derivação que é feita a verificação da adequação das sequências de palavras às regras de construção impostas pela linguagem, na composição de frases, períodos ou orações.

A figura 2 ilustra a árvore de derivação criada como resultado da análise sintática da frase “Eu quero imprimir o arquivo .init do Mário.”.

Na análise sintática testa se os sintagmas foram postos na sequência correta, ou seja, se por exemplo, dois substantivos podem se seguir ou não (SCHNEIDER, 2001). Uma das principais técnicas para realizar a análise sintática é conhecida etiquetagem *outagging*.

O processo de análise sintática pode ser chamado de *parsing*.

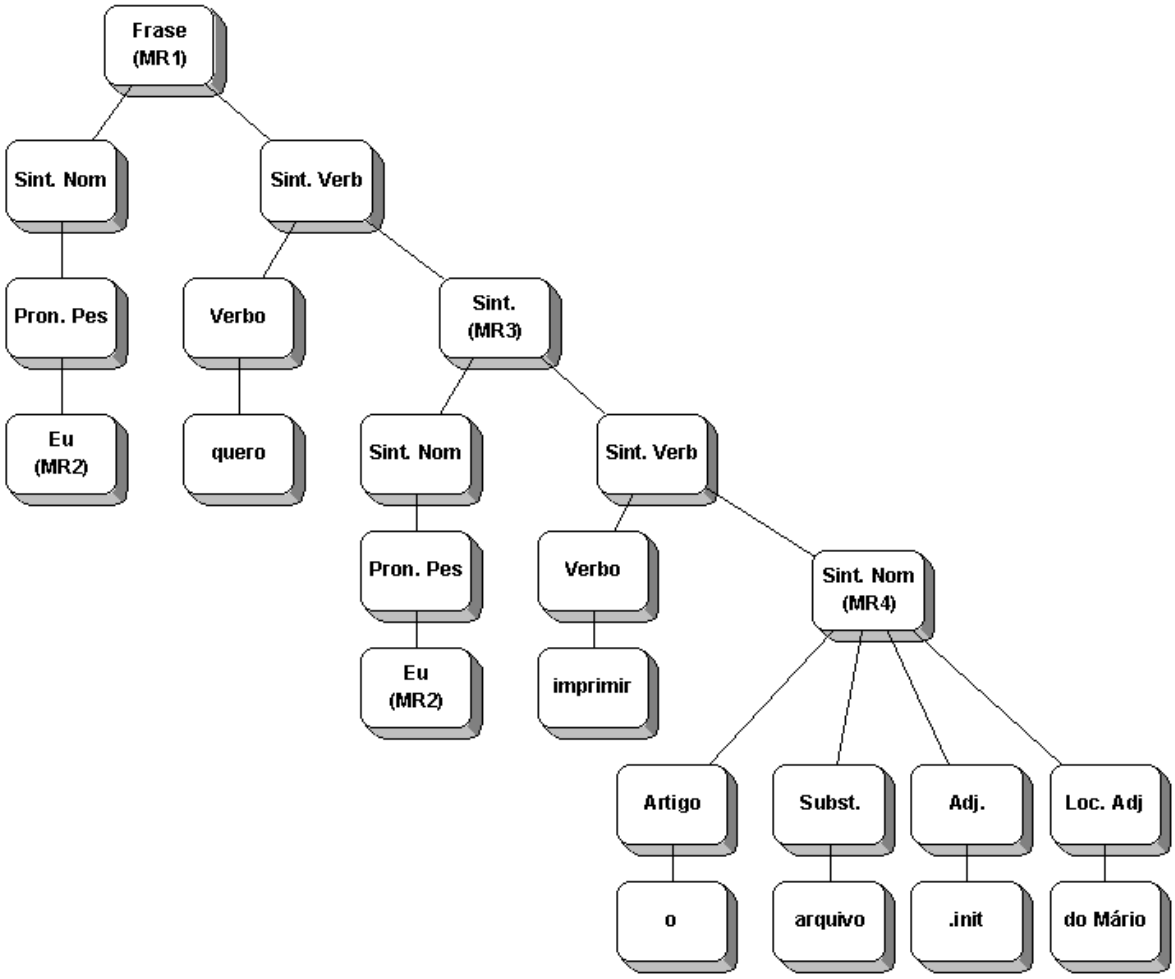

Figura 3 - Resultado da análise sintática. Fonte: OLIVEIRA (2011)

2.1.2.3.1 Etiquetagem

Na Linguística Computacional, etiquetagem consiste na atribuição de categorias a porções do texto (OLIVEIRA et al., 2006), o a ferramenta responsável por esse processo é conhecida como etiquetador e normalmente é desenvolvida para um *parser* específico.

Basicamente a etiquetagem dos *parsers* atualmente implantados constitui-se de três tipos: os que utilizam regras, os que utilizam estatísticas e os que utilizam os dois que são conhecidos como híbridos (JURAFSKY et al., 2000).

Os *parsers* que utilizam as regras para efetuarem a etiquetagem classificam as palavras consultando seu dicionário de palavras devidamente classificadas, esse dicionário é chamado de ontologia.

2.1.2.4 Análise Semântica

Conforme OLIVEIRA (2011), a questão da representação do conhecimento apresenta diversos problemas. Podemos citar a o problema da ambiguidade como tomar em “tomar um banho” e em “tomar de alguém” ou até mesmo em “tomar um copo de suco”, além de outros.

OLIVEIRA (2011) também classifica a semântica em léxica e em gramatical, a semântica léxica busca descrever o sentido através do uso de da decomposição semântica das unidades léxicas ou através das redes semânticas que leva em conta como os humanos memorizam e categorizam os conceitos. Já a semântica gramatical tenta buscar o sentido através de uma formula lógico-semântica, porém há casos que em uma estrutura pode dar origem a duas representações semânticas, como em “uma professora de capoeira pernambucana” pode referir-se a uma pessoa nascida em Pernambuco ou a uma pessoa que ensina capoeira no estilo Pernambucano.

2.1.2.5 Análise Pragmática

Tal etapa consiste em reinterpretar frases de forma que seu significado real seja definido. Uma vez que em frases como a pergunta: "Você sabe que horas são?" o locutor deseja saber que horas são naquele momento e não apenas saber se o ouvinte sabe que horas são. Esta análise é aplicada apenas a situações específicas em que há ambiguidade no seu sentido.

A análise pragmática não se restringe apenas a uma frase, ela busca o significado no contexto em que a frase se encontra.

2.2 MÉTODOS DE REPRESENTAÇÃO DO CONHECIMENTO

Os sistemas de Recuperação da Informação utilizam-se de dicionários para consulta, esses dicionários podem ser estruturados de várias formas dependendo de sua utilidade.

2.2.1 Tesauro

O Moderno Dicionário da Língua Portuguesa Michaelis define Tesauro como uma coleção de palavras agrupadas por conceitos e títulos, e não em ordem alfabética como num dicionário ou para o domínio da informática como arquivo contendo sinônimos que são exibidos como alternativas para uma palavra escrita de forma incorreta, durante uma verificação de ortografia.

O termo Tesauro também pode ser comumente encontrado em sua forma no latim, *thesaurus.*

2.2.2 Corpus

Na concepção de DUCROT e TODOROV (2001), *corpus* é um "conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida linguagem determinada época". Para TRASK (2004) *apud* ALOÍSIO e ALMEIDA (2006), *corpus* é "um conjunto de textos escritos ou falados numa língua, disponível para análise".

2.2.3 Ontologia

GRUBER (2009) define, em tradução livre, ontologia no contexto da ciência da computação e informação como um conjunto de representações primitivas de um domínio de conhecimento ou discurso específico. A representação primitiva é tipicamente realizada através de classes (ou conjuntos), atributos (ou propriedades) e relacionamentos[1].

[1] In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships.

2.3 RECUPERAÇÃO DA INFORMAÇÃO

Conhecido por RI ou *information retrieval* (IR), a recuperação da informação surgiu em 1950 e tem como meta encontrar a informação exigida para satisfazer a necessidade de informação (NI) do usuário (FRANTZ 1997 apud GONZALEZ et al. 2011).

Recuperação de informação é uma subárea da ciência da computação que estuda o armazenamento e recuperação automática de documentos, que são objetos de dados, geralmente textos. (CARDOSO, 2011)

Segundo TEIXEIRA e SCHIEL (1997), o processo de recuperação da informação compreende basicamente três etapas: indexar, armazenar e recuperar. Com o avanço tecnológico essas tarefas ficaram mais fácies de serem realizadas.

2.3.1 Sistemas de Recuperação da Informação

São Sistemas de Recuperação da Informação quem realmente realizam todo o trabalho de RI, eles consultam um banco de dados atrás da informação requisitada e filtram os resultados obtidos em busca da melhor resposta.

GONZALEZ et al. (2011) divide os Sistemas de Recuperação da Informação em três componentes iniciais: o usuário, a necessidade de informação do usuário e a coleção de documentos disponíveis para pesquisa. Sem esses componentes é impossível realizar a RI.

Existem componentes adicionais que são: a consulta, que traduz a necessidade de informação do usuário, os índices dos documentos para auxiliar e otimizar a consulta e a referência de cada elemento da coleção, a referência é normalmente chamada de *surrogate* e pode ser composta de título, autor, data, resumo, etc.

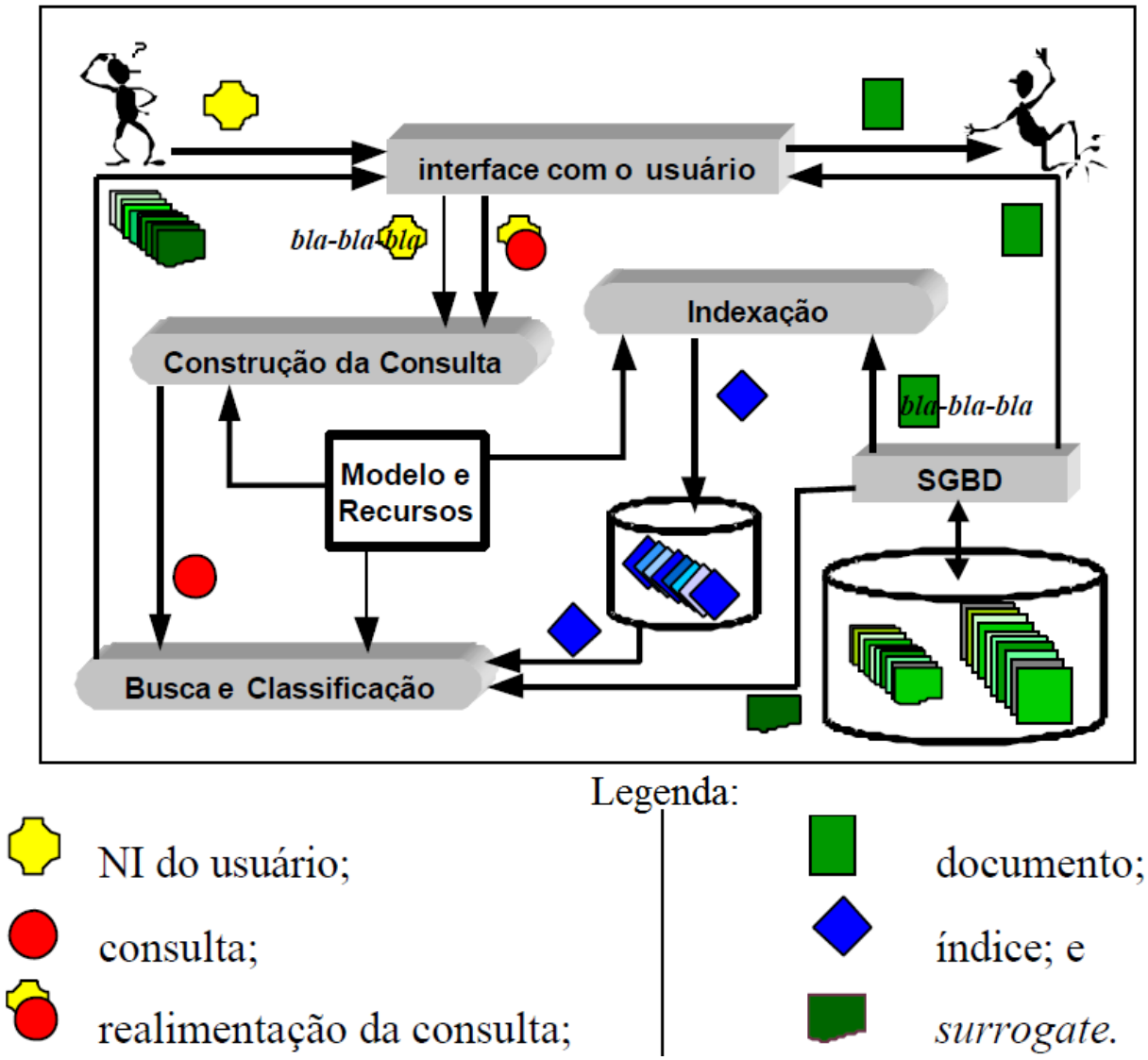

Figura 4 - Esquema geral de um sistema de RI. Fonte: Gonzalez et al. (2011)

2.3.2 Sistemas de Recuperação da Informação e Ontologias

As ontologias são consultadas pelos Sistemas de RI em busca de respostas para alguma pergunta. A Figura 4 ilustra como é feito a comunicação Sistema de RI/ontologias.

Através de uma ontologia pode definir-se uma hierarquia de classes representativas de conceitos, objetos ou entidades, caracterizadas pelas respectivas propriedades (SAIAS, 2003).

2.4 SISTEMAS DE PERGUNTAS E RESPOSTAS

Os Sistemas de Perguntas e Respostas ou Q&A, do inglês *Question and Answer,* podem ser definidos como sistemas que utilizam algum tipo de linguagem (natural ou de máquina) como meio de comunicação que recebe uma pergunta que é processada através de um *parser e* fornece uma resposta consultado seu *corpus* ou tesauro.

Alguns sistemas de Q&A utilizam um *corpus* finito para responder as perguntas, obtendo-se muitas vezes resultados mais precisos, outros utilizam *corpus* sem tamanho definido, como por exemplo, a Internet.

Um dos problemas de se utilizar a Internet como *corpus* é o fato de que os resultados poderão ser menos precisos e menos confiáveis do que os de tamanho finito, porém possuem um domínio de conhecimento do tamanho da Internet.

Os tipos de perguntas e de respostas podem variar dependendo do tipo, do domínio e do objetivo desse sistema.

2.4.1 Arquitetura dos Sistemas de Perguntas e Respostas

Segundo QUARESMA et al. (2006), os sistemas de perguntas e respostas possuem dois módulos, um responsáveis pelo processamento do corpus e outro pelo processamento da pergunta. A Figura 5 ilustra como é a arquitetura desse tipo de sistema, que foi a utilizada para a construção do protótipo e o processo que ocorre na etapa de recuperação da informação está ilustrado na Figura 4.

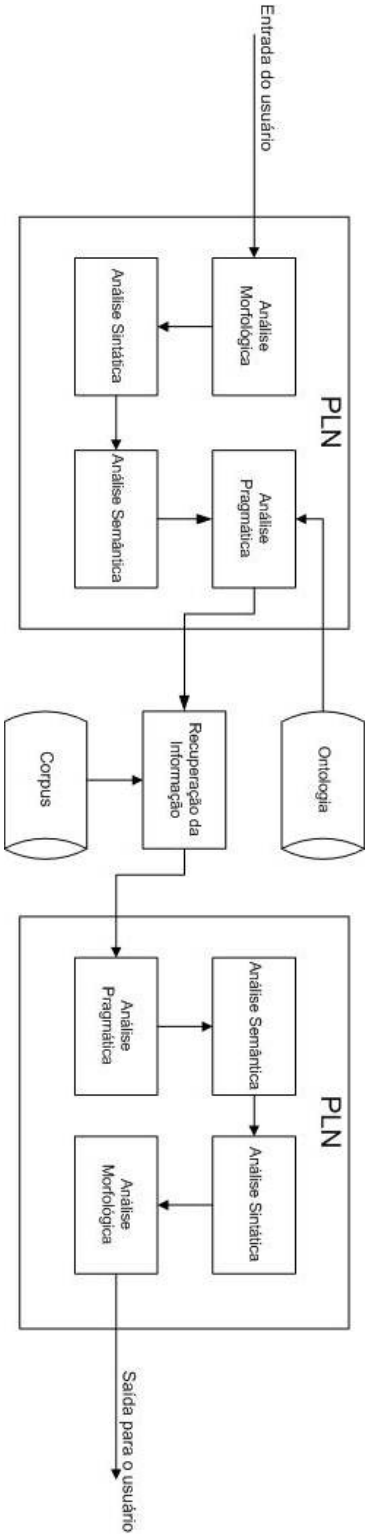

Figura 5 - Arquitetura de um sistema de perguntas e resposta

3 ANÁLISE DAS FERRAMENTAS

Para construção do protótipo foram selecionadas e testadas algumas ferramentas que são utilizadas para realização de etapas do Processamento de Linguagem Natural e alguns Sistemas de Perguntas e Respostas para analise de seu funcionamento e arquitetura.

3.1.1 Análise dos *parsers*

Os *parsers* encontrados foram submetidos à testes no dia 09 de maio de 2011, utilizando-se a frase “O ministro de Minas e Energia, Edison Lobão, disse que a partir desta segunda-feira o preço dos combustíveis nos postos de gasolina vai cair.” retirada da seção de economia do site Universo Online (UOL)[2] no dia 09 de maio de 2011, e as principais dificuldades podem ser observadas no Quadro 1.

Quadro 1 - Quadro comparativo dos *parsers.*

Ferramenta	Dificuldades
Curupira – O *parser* para o português brasileiro	• Não está disponível ao público.
Grammar Play	• Anotação manual: frase por frase; • Aceita somente pequenas frases; e • Reconhece somente verbos conjugados na terceira pessoa do singular ou plural.
The Stanford Parser: A statical parser	• Feito para outro idioma; • Necessita de dicionário auxiliar; e • Problemas de acentuação.
Multilingua	• Funciona somente no Linux.
DepPattern	• Funciona somente no Linux.
Unitex	• Difícil integração com outros sistemas.

[2] http://economia.uol.com.br/ultimas-noticias/redacao/2011/05/09/ministro-diz-que-a-partir-de-hoje-consumidor-deve-pagar-menos-por-combustiveis.jhtml

3.1.1.1 Stanford *Parser*

Desenvolvido pelo grupo de estudos *The Stanford Natural Language Processing Group* da Universidade de Stanford, o Stanford Parser é desenvolvido utilizando-se a linguagem de programação Java e tem como foco o idioma Inglês. Para que ele funcione em outros idiomas é necessário a utilização de bibliotecas adicionais. Oficialmente existem bibliotecas para os idiomas Inglês, Alemão, Chinês, Arábico e Frances (STANFORD, 2011).

A biblioteca para o Português foi desenvolvida pelo *Language Resources and Technology for Portuguese* da Universidade de Lisboa (SILVA et al. 2010).

Sua principal vantagem é que seu código fonte é livre e pode ser integrado ao código fonte do protótipo não havendo a necessidade de que o protótipo tenha que analisar o arquivo texto gerado pelo parser e sua principal desvantagem é que há a necessidade de utilizar uma biblioteca auxiliar para análise do Português.

Como pode ser observado na Figura 6, o parser analisa a entrada do usuário e retorna o texto com a análise em forma de árvore, onde cada elemento possui a palavra e sua classificação, como por exemplo a palavra “de” precede foi classificada como uma preposição (P) e “disse” como verbo (V).

```
F:\_Arquivos\TCC\parser>java -Xmx500m -cp .;stanford-parser.jar edu.stanford.nlp
.parser.lexparser.LexicalizedParser -tokenized -sentences newline -outputFormat
oneline -uwModel edu.stanford.nlp.parser.lexparser.BaseUnknownWordModel cintil.s
er.gz "F:\_Arquivos\TCC\data\portuguese.onesentece.txt"
Loading parser from serialized file cintil.ser.gz ... done [0,9 sec].
F:\_Arquivos\TCC\data\portuguese.onesentece.txt
Parsing file: F:\_Arquivos\TCC\data\portuguese.onesentece.txt with 1 sentences.
Parsing [sent. 1 len. 24]: [O, ministro, de, Minas, e, Energia,, Edison, Lobao,,
 disse, que, a, partir, desta, segunda-feira, o, preco, dos, combustiveis, nos,
postos, de, gasolina, vai, cair.]
(ROOT (S (S (NP (NP (D O) (N' (N ministro) (PP (P de) (NP (N Minas))))) (NP (CON
J e) (NP (N' (N Energia,) (N Edison) (N Lobao,))))) (VP (V disse) (CP (C que) (S
 (NP (D a) (N' (N partir) (N desta))) (VP (V segunda-feira) (NP (D o) (N' (N' (N
 preco) (N dos) (N combustiveis) (N nos) (N postos)) (PP (P de) (NP (N gasolina)
)))))))) (S (V vai) (NP (N cair.)))))
Parsed file: F:\_Arquivos\TCC\data\portuguese.onesentece.txt [1 sentences].
Parsed 24 words in 1 sentences (13,17 wds/sec; 0,55 sents/sec).
```

Figura 6 - Funcionamento do Stanford Parser

3.1.1.2Multilingua

O parser Multilingua foi desenvolvido pela Universidade de Santiago da Compostela e atualmente funciona para os idiomas Inglês, Espanhol, Galego, Frances

e Português. Foi implementado em Perl e seus desenvolvedores disponibilizam uma distribuição que funciona somente em uma distribuição do sistema operacional Linux (USC, 2008).

Após análise do parser Multilingua foi possível concluir que mesmo sendo feito para o idioma Português apresenta algumas limitações, como ser limitado à distribuições Linux e o resultado ser sempre em arquivo texto.

O texto abaixo ilustra o resultado obtido utilizando o parser Multilingua, que assim como o Stanford Parser utiliza sua própria anotação, no Multilingua a classificação vem após a palavra, como pode ser visto a palavra “dizer” veio acompanhada de _VERB que é a nomenclatura que classifica a palavra como verbo.

```
SENT::<o_DT_0 ministro_NOM_1 de_PRPP_2 Minas_NOM_3 e_CONJ_4 energia_NOM_5
,_VERBF_6 Edison&Lobão_NOM_7 ,_VERBF_8 dizer_VERBF_9 que_CS_10 a_PN_11
partir_VERBI_12 de_PRPP_13 segunda-feira_NOM_14 o_DT_15 preço_NOM_16
de_PRPP_17 combustível_NOM_18 em_PRP_19 posto_NOM_20 de_PRPP_21
gasolina_NOM_22 ir_VERBF_23 cair_VERBI_24 ._SENT>
(Spec;ministro_NOM_1;o_DT_0)
(Comp;de_PRPP_2;Minas_NOM_3)
(de_PRPP_2;ministro_NOM_1;Minas_NOM_3)
(Lobj;,_VERBF_6;energia_NOM_5)
(Robj;,_VERBF_6;Edison&Lobão_NOM_7)
(Lobj;partir_VERBI_12;a_PN_11)
(Robj;dizer_VERBF_9;partir_VERBI_12)
(Spec;preço_NOM_16;o_DT_15)
(Robj;partir_VERBI_12;preço_NOM_16)       (Comp;de_PRPP_17;combustível_NOM_18)
(de_PRPP_17;preço_NOM_16;combustível_NOM_18) (Comp;em_PRP_19;posto_NOM_20)
(em_PRP_19;preço_NOM_16;posto_NOM_20)
(de_PRPP_21;posto_NOM_20;gasolina_NOM_22) (Comp;de_PRPP_21;gasolina_NOM_22)
(Spec;cair_VERBI_24;ir_VERBF_23)
---
```

3.1.1.3 DepPattern

O DepPattern é um pacote linguístico que provê um compilador gramatical, um PoS tagger (etiquetador) e um analisador de dependências. O parser oferece suporte a cinco idiomas, sendo eles Inglês, Espanhol, Galego, Francês e Português. Assim como o Multilingua foi desenvolvido pela Universidade de Santiago de Compostela e também só funciona em uma distribuição Linux (USC, 2011).

O DepPatter apresentou o melhor análise gramatical de todos os parsers estudados, porém possui as mesmas limitações encontradas no parser Multilingua e o tempo para processamento foi de quase um minuto, sendo assim, maior que todos estudados.

O texto a seguir ilustra um trecho o funcionamento do DepPattern e como pode ser observado seu resultado é mais detalhado pois mostra o gênero, origem, modo, pessoa, número da palavra e se possui correspondência a outra palavra.

```
root@Lab:/home/lab/Download/DepPattern-2.0# sh dp.sh -fatreetaggerpt texto.txt
SENT::O_DT_0_<gender:0|lemma:o|number:0|person:0|pos:0|possessor:0|token:O|typ
e:0|>
ministro_NOUN_1_<gender:0|lemma:ministro|number:0|person:3|pos:1|token:ministr
o|type:C|>
de_PRP_2_<lemma:de|pos:2|token:de|type:0|>
Minas_NOUN_3_<gender:0|lemma:Minas|number:0|person:3|pos:3|token:Minas|type:P|
>
e_C_4_<lemma:e|pos:4|token:e|type:0|>
Energia_NOUN_5_<gender:0|lemma:energia|number:0|person:3|pos:5|token:Energia|t
ype:C|>                                          ,_Fc_6_<lemma:,|pos:6|token:,|>
Edison&Lobão_NOUN_7_<gender:0|lemma:Edison&Lobão|number:0|person:3|pos:7|token
:Edison&Lobão|type:P|>
,_Fc_8_<lemma:,|pos:8|token:,|>
disse_VERB_9_<gender:0|lemma:dizer|mode:0|number:0|person:0|pos:9|tense:0|toke
n:disse|type:0|>
que_C_10_<lemma:que|pos:10|token:que|type:S|>
a_PRO_11_<case:0|gender:0|lemma:a|number:0|person:0|politeness:0|pos:11|posses
sor:0|token:a|type:0|>
```

3.2 SISTEMAS DE PERGUNTAS E RESPOSTAS

A fim de se entender melhor o seu funcionamento e sua arquitetura alguns Sistemas de Perguntas e Respostas foram analisados e o resultados podem ser observados a seguir.

3.2.1 NJFun

De acordo com LITMAN et al. (2000), o NJFun é um sistema de pergunta e resposta para o idioma Inglês que tem como objetivo principal oferecer lugares e atrações no estado de New Jersey, Estados Unidos.

Seu banco de dados foi populado a partir do nj.online, principal site com informações sobre o estado e foi indexado utilizando-se três atributos principais: tipo de atividade, localização e período do dia (que pode assumir manhã, tarde e noite).

A extração das informações com o usuário é feito através de várias perguntas formuladas de maneira diferente para se obter um dos três atributos, feito isso o processo se repete até que todos estejam completos. Após as perguntas o sistema efetua uma consulta no banco de dados utilizando o *wildcard* para os atributos que não foram preenchidos.

O NJFun realiza de uma a 12 perguntas até consultar no banco de dados.

3.2.2 Deal

O sistema Deal tem como foco o aprendizado de idiomas. Seu funcionamento é feito através de uma interface gráfica onde o aluno tem que comprar objetos de um personagem (*non-player characters, NPC*) e pechinchar no preço.

Apesar de possuir um número limitado de “produtos à venda” a recuperação da informação sobre os produtos é feita através da análise de palavras-chave extraídas por meio do PLN (HJALMARSSON et al., 2007).

3.2.3 Conclusões das análises

Após as analises das ferramentas de perguntas e respostas foi possível observar que a extração das perguntas são feitas sempre por meio de padrões e que para que o sistema possa fazer a recuperação da informação em seu Corpus ele precisa de conjunto mínimo de informações de tamanho variável de sistema para sistema.

4 APLICAÇÃO

Após análise e estudo das ferramentas foram selecionados um parser, um etiquetador e um sistema de gerenciamento de banco de dados para a construção do protótipo.

4.1 O PROTÓTIPO DO SISTEMA

Existem diversas ferramentas que realizam uma determinada etapa do processamento de linguagem natural, e até mesmo sistemas que realizam a mesma tarefa do sistema proposto, porém nenhum que funcione efetivamente para o idioma Português Brasileiro. A ferramenta que faz a análise do texto em Português enviado pelo utilizador ao computador transformando-o e interpretando-o gerando, assim, uma resposta através da recuperação da informação em um banco de dados.

A resposta ao utilizador também deverá ser em português e será gerada por meio do processo inverso da pergunta.

Para tanto foi desenvolvido uma integração das ferramentas já existentes que foram selecionadas após análise e o desenvolvimento de um interpretador que entenda o resultado do PLN e faça a recuperação da informação em seu Corpus.

O domínio escolhido para desenvolvimento do protótipo foi de um sistema para consulta de horários de ônibus.

4.2 DESENVOLVIMENTO DO PROTÓTIPO

Após estudos foi possível detectar que a maior parte de ferramentas envolvidas nas etapas do Processamento de Linguagem Natural eram desenvolvidas nas linguagens de programação Java e Perl, porém após testes com estas ferramentas observou-se que as ferramentas desenvolvidas na linguagem Java possuíram resultados mais satisfatórios além de apresentarem mais facilidade de integração. Portanto a linguagem de programação utilizada no protótipo escolhida foi a linguagem Java.

O protótipo inicia sempre requisitando ao usuário que faça sua pergunta, que deverá estar com sua gramática parcialmente correta (nomes próprios com a letra inicial maiúscula, ponto de interrogação no final da frase, etc.) e sempre terá que possuir a origem e o destino do usuário antes de fazer qualquer recuperação da informação. Como pode ser observado na Figura 7, se o usuário não informar essas informações o protótipo fará perguntas até que a consiga.

```
Bem vindo, o que voce deseja saber?
Quais são os horários de ônibus para Bandeirantes?
Loading parser from serialized file grammar/cintil.ser.gz ... done [0.8 sec].
Não entendi a origem, por favor, digite-a novamente:
Londrina
Encontrei o seguintes resultados:
+------------------+------------------+------------------+------------------+------------------+
|     EMPRESA      |       ORIGEM     |       DESTINO    |    HORARIO SAIDA |  HORARIO CHEGADA |
+------------------+------------------+------------------+------------------+------------------+
| GARCIA           | LONDRINA         | BANDEIRANTES     | 06:30:00         | 08:30:00         |
| OURO BRANCO      | LONDRINA         | BANDEIRANTES     | 08:00:00         | 10:05:00         |
| GARCIA           | LONDRINA         | BANDEIRANTES     | 09:30:00         | 11:30:00         |
| GARCIA           | LONDRINA         | BANDEIRANTES     | 11:00:00         | 13:00:00         |
| OURO BRANCO      | LONDRINA         | BANDEIRANTES     | 12:40:00         | 14:45:00         |
| GARCIA           | LONDRINA         | BANDEIRANTES     | 14:00:00         | 16:00:00         |
| OURO BRANCO      | LONDRINA         | BANDEIRANTES     | 15:50:00         | 17:50:00         |
| GARCIA           | LONDRINA         | BANDEIRANTES     | 17:30:00         | 19:30:00         |
| OURO BRANCO      | LONDRINA         | BANDEIRANTES     | 19:00:00         | 21:05:00         |
| GARCIA           | LONDRINA         | BANDEIRANTES     | 21:00:00         | 23:00:00         |
+------------------+------------------+------------------+------------------+------------------+
```

Figura 7 – Exemplo de funcionamento do protótipo.

4.2.1 Parser

O parser que apresentou melhor desempenho e facilidade de utilização e integração foi o Stanford Parser com a adaptação para o reconhecimento e análise de textos no idioma Português desenvolvido pela Universidade de Lisboa.

4.2.2 Etiquetador

O etiquetador escolhido para pré-processamento do texto foi o Stanford POS Tagger (TOUTANOVA et al., 2003) por possuir excelente compatibilidade com o Stanford Parser.

4.2.3 Padrões de Pergunta

A interpretação das perguntas pelo protótipo foi feita por meio de busca e análise de padrões de pergunta.

A informação necessária é extraída utilizando-se os padrões de pergunta e o resultado do parser, por exemplo, a preposição “de” seguida de um nome próprio tem alta probabilidade de identificar a origem do usuário, assim como a palavra “Qual” define que ele quer saber somente um horário. A Figura 8 mostra como é o funcionamento para extração da informação, quando o usuário digita “Quais são os horários de ônibus para Londrina?” o protótipo identifica que se trata de uma pergunta requisitando um horário de ônibus, ou seja, deverá recuperar a informação na tabela horários e o destino requisitado é Londrina serve como filtro para refinar a pergunta.

Figura 8 – Funcionamento para extração de resposta.

No total foram construídos 32 padrões de extração de resposta, o Quadro 2 ilustra alguns modelos de perguntas e quais informações podem ser extraídas através daquele padrão.

Quadro 2. Padrões de perguntas e informações extraídas.

Pergunta	Informações
Quais são os horários de ônibus de {Cidade} para {Cidade}?	Origem, destino e o usuário deseja ver vários horários.
Qual o próximo ônibus de {Cidade} para {Cidade}?	Origem, destino e o usuário deseja ver somente o próximo horário.
Tem algum ônibus que parte de {Cidade} para {Cidade} pela {Empresa}?	Origem, destino e empresa de transporte.

4.2.4 Resposta

A resposta é dada sempre em forma de quadro, quando o usuário requisita por múltiplas informações ou em forma de texto quando requisita somente uma informação, a resposta em forma de quadro contendo a empresa de transporte, a origem, o destino, o horário de saída e o horário previsto de chegada, como pode ser vista na Figura 8.

4.2.5 Recuperação da Informação

O banco de dados utilizado foi montado de forma manual através de informações contidas na internet e possui a empresa de transporte responsável pela viagem, o destino, a origem, o horário de saída e o horário provável de chegada. O Sistema Gerenciador de Banco de Dados escolhido foi o MySQL, por ser grátis, leve e rápido.

5 RESULTADOS

O protótipo foi testado durante o período de 07 a 18 de novembro de 2011 por 15 pessoas, com idade de 20 a 60 anos possuindo ensino variando entre médio e superior, cada uma realizou em média de duas a três perguntas ao sistema. Foram feitas no total 40 perguntas sendo que dessas 29 tiveram um resultado satisfatório. Considerou-se a resposta satisfatória, aquela que consegue passar a informação desejada pelo usuário. Em apenas 11 perguntas o sistema não conseguiu formular uma resposta correta, sendo assim, conforme pode ser analisado no gráfico da Figura 9, a porcentagem de respostas corretas foi de 72% contra 28% de respostas incorretas.

Alguns exemplos das perguntas realizadas pelos testadores foram:

1. Quais são os horários de ônibus de Londrina para Bauru?
2. Quais ônibus partem de Londrina para Bandeirantes?
3. Qual o próximo ônibus de Cornélio Procópio para Bandeirantes?
4. Qual empresa que parte de Londrina?
5. Quais ônibus partem às 14:00 horas?
6. Quais são os próximos 10 ônibus que partem de Andirá?
7. Quais empresas possuem ônibus que saem às 10:00 de Avaré?
8. Quais ônibus chegam às 14:00 horas?
9. Existe algum ônibus de Avaré para Arandu?
10. Há algum ônibus de Araraquara para Bauru?
11. Qual o próximo ônibus de São Paulo para Londrina?
12. Que horas sai o ônibus de Londrina para Ourinhos?
13. Tem algum horário de Londrina para Bandeirantes?
14. Quais empresas possuem ônibus que saem às 10:00 de Londrina para Bandeirantes?
15. Qual empresa tem ônibus que parte de Florianópolis?

Figura 9 – Resultado das respostas.

Os erros foram analisados e os fatores determinantes dos erros foram descritos na Tabela 1 e conforme ilustra a Figura 10 o principal motivo com 46% foi o erro ao interpretar a pergunta feita pelo usuário.

Quadro 3 – Motivos que levaram à respostas incorretas

Motivo	Número de ocorrências
Erro de interpretação da pergunta	5
Erro ao recuperar a informação	2
Erro ao formular a resposta	2
Erro ao exibir a resposta	2

Figura 10 – Motivos dos erros.

Os erros de interpretação da pergunta foram gerados sempre na etapa de *parsing* ou na etiquetagem, sendo possível concluir que a biblioteca utilizada para adaptação do Stanford Parser ao Português ainda possui erros. Os erros ao recuperar a informação ocorreram quando o protótipo não conseguiu construir com sucesso a consulta ao banco de dados.

Foram considerados erros de formulação da resposta aqueles em que o protótipo exibiu resultados que não foram requisitados pelo usuário, como por exemplo exibir horários de ônibus para outra cidade e os erros ao exibir a resposta quando o protótipo exibiu um número errado de resultados, como por exemplo exibir mais de um resultado quando o usuário requisitou somente um horário.

Abaixo estão relacionadas algumas perguntas que resultaram em erro:

1. Estou em Avaré e quero ir para Londrina, qual é o próximo ônibus?
2. Preciso chegar em Bauru o mais rápido possível, qual é o próximo?
3. Tem algum ônibus para o Paraguai?
4. Quais são os ônibus para eu ir para Curitiba?
5. Que horas sai o próximo ônibus para Bandeirantes?

6 CONCLUSÕES

O trabalho realizado mostrou e proporcionou uma possível solução para diminuir as dificuldades da comunicação homem/máquina por meio do Processamento de Linguagem Natural e da Recuperação da Informação desenvolvendo um protótipo de um Sistema de Perguntas e Respostas em que o usuário não precisa utilizar-se de uma linguagem de programação para realizar essa comunicação, evitando-se assim a necessidade do aprendizado de uma nova linguagem pelo usuário.

Para tanto foi necessário o estudo sobre o Processamento de Linguagem Natural e uma pesquisa por ferramentas que realizam cada uma de suas etapas para que o protótipo identifique o que o usuário está escrevendo. Estudou-se também Recuperação da Informação para entender como o protótipo deveria proceder para recuperar a informação requisitada no banco de dados e sobre Sistemas de Perguntas e Respostas para definir a arquitetura do protótipo e como deveria ser seu funcionamento.

Os *parsers* encontrados foram submetidos à testes para analisar a sua velocidade de processamento, se o resultado obtido estava correto e sua capacidade de integração com o protótipo.

O desenvolvimento do protótipo foi feito por meio da integração das ferramentas escolhidas e do desenvolvimento de um módulo que entenda a saída do *parser*, recupere a informação em um banco de dados e forneça uma resposta ao usuário.

A arquitetura definida para o desenvolvimento foi definido após analisar e estudar o funcionamento de outros Sistemas de Perguntas e Respostas semelhantes ao protótipo proposto.

Após a construção de um banco de dados com informações retiradas da internet o protótipo foi submetido a testes e foi possível observar que ele conseguiu responder a maior parte das perguntas e que ainda é possível melhorar seu desempenho e incrementá-lo.

Portanto pode-se concluir que apesar de possuir alguns erros, o protótipo apresentou um resultado satisfatório e mostrou que o Processamento de Linguagem Natural pode auxiliar no processo de Recuperação da Informação. Os testes e o seu uso mostrou que ele é útil e que ainda há muito coisa que possa ser feito para incrementá-lo, como a criação de uma interface gráfica e síntese da voz.

7 TRABALHOS FUTUROS

Ainda há muito o que ser feito e estudado dentro da área do Processamento de Linguagem Natural, da Recuperação da Informação e de aperfeiçoamento do protótipo proposto, como por exemplo:

- Criação de uma interface gráfica (GUI);
- Análise e síntese da voz;
- Melhor adaptação ao Português;
- Correção de bugs e otimização; e
- Recuperação da Informação da internet e não de um banco de dados local.

REFERÊNCIAS

ALMEIDA, S.; CARVALHO, A.; FANTIN, L.; STOLFI, J. Selva: *A New Syntatic Parser for Portuguese.*

ALUÍSIO, S. M.; ALMEIDA, G. M. de B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa lingüística. **Calidoscópio**, Vol. 4. M. 3, p. 155-177, set/dez 2006.

BIRD, S; KLEIN, E.; LOPER, E. In: Natural Language Processing with Python. O´REILLY, 1ª ed.,2009.

CARDOSO, O. N. P. **Recuperação de Informação**. Universidade Federal de Lavras. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf>. Acesso em 17 agosto 2011.

DUCROT, O. e TODOROV, T. In: Dicionário enciclopédico das ciências da linguagem.Perspectiva, 3ª ed., 2001, São Paulo, p. 339.

FRANTZ, V.; SHAPIRO, J.;VOISKUNSKII, V. *Automated Information Retrieval: Theory and Methods*. **Academic Press**. 1997 p. 365.

GONZALEZ M.; LIMA, V. L. S. **Recuperação da Informação e Processamento de Linguagem Natural.** Porto Alegre, RS. Disponível em: <http://www.inf.pucrs.br/~gonzalez/docs/minicurso-jaia2003.pdf>. Acesso em 11 agosto 2011.

GONZALEZ M.; LIMA, V. L. S.; LIMA, J. V. Tools for Nominalization: an Alternative for Lexical Normalization. ***Workshop on Computional Procedure of the Portuguese Language****.* 2006 p. 100-109.

GRUBER, T. Ontology. **Encyclopedia of Database Systems**, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009. Disponível em: <http://tomgruber.org/writing/ontology-definition-2007.htm>. Acesso em: 07 maio 2011.

HJALMARSSON, A.; WIK, P.; BRUSK, J. *Dealing with Deal: A dialogue system for conversation training.* ***Proceedings of the 8th SIGdial Workshop Discourse and Dialogue.*** 2007 p. 132-135.

JURAFSKY, D.; MARTIN, J.H. In: **Speech and Language Processing** - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000.

KLEIN, D.; MANNING, C. D.; *Accurate Unlexicalized Parsing.* ***Proccedings of the 41st Meeting of the Association for Computational Linguistics***, p 423-430.

LINGUATEC. **Linguatec Sprachtechnologien GmbH:** Grundlagen der Spracherkennung. 2001. Disponível em*:* <http://www.spracherkennung.de/service/sebuch.htm >. Acesso em: 05 maio 2011.

LITMAN, D. J.; KEARNS, M. S.; SINGH, S.; WALKER, M. A. *Automatic Optimization of Dialogue Management*. ***Proccedings of the 18th conference on Computational Linguistics.*** Alemanha, p 502-508, agosto 2000.

MARTINS, R. T.; HASEGAWA, R.; NUNES, M. G. V. **Curupira: um parser funcional para o português.** 2002. Disponível em < http://www.nilc.icmc.usp.br/nilc/download/nilc-tr-02-26.zip> . Acesso em: 06 maio 2011.

MENDES, A. J. N. **Software educativo para apoio à aprendizagem de programação.** Universidade de Coimbra. Portugal. Disponível em: http://www.c5.cl/ieinvestiga/actas/tise01/pags/charlas/charla_mendes.htm. Acesso em: 11 agosto 2011.

Moderno Dicionário da Língua Portuguesa Michaelis. Editora Moderna, 2011.

MÜLLER, D. N. **Processamento de Linguagem Natural.** Universidade Federal do Rio Grande do Sul. Porto Alegre / RS. 2003. Disponível em: <http://www.inf.ufrgs.br/~danielnm/docs/daniel-nehme-muller-tese.pdf>. Acesso em: 06 maio 2011.

OLIVEIRA, C.; FREITAS, M. C. Classe de palavras e etiquetagem na Lingüística Computacional. **Calidoscópio**, Vol. 4, n. 3, p179-188, set/dez 2006.

OLIVEIRA, F. A. D. **Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa.** Universidade Federal do Rio Grande do Sul. Porto Alegre/RS. Disponível em: <http://www.inf.ufrgs.br/gppd/disc/cmp135/trabs/992/Parser/parser.html>. Acesso em 01 junho 2011.

OTHERO, G. A. ***Grammar Play:*** **um parser sintátito em Prolog para a lingual portuguesa.** Pontifícia Universidade do Rio Grande do Sul, Porto Alegre/RS. 2004.

QUARESMA, P.; RODRIGUES, I.; PROLO, C. A.; VIEIRA, R. Um sistema de Pergunta-Resposta para uma base de Documentos. **Letras de Hoje**. Porto Alegre. v. 41, nº 2, p. 43-63, junho, 2006.

RICH, E.; KNIGHT, K. Inteligência Artificial. **Makron Books**, 1993, 722p.

ROSA, J. L. G. O Processamento de Linguagem Natural. Diário do Povo, Campinas/SP, 14/09/1995. Caderno de Informática.

SAIAS, J. M. G. **Uma Metodologia para a construção automática de Ontologias e sua aplicação em Sistemas de Recuperação da Informação**. Universidade de Évora. 2003.

SCHNEIDER, M. O. **Processamento de Linguagem Natural (PLN).** Pontifícia Universidade Católica de Campinas. Campinas / SP. 2001. Disponível em: <http://moschneider.tripod.com/pln.pdf>. Acesso em: 05 maio 2011.

SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R. *Out-of-the-Box Robust Parsing of Portuguese.* ***Proceedings of the 9th International Conference on the Computational Processing of Portuguese.*** 2010 p. 75-85.

STANFORD. ***The Stanford Parser: A Statistical parser***. Disponível em: < http://nlp.stanford.edu/software/lex-parser.shtml>. Acesso em: 06 julho 2011.

TEIXEIRA, C. M. S.; SCHIEL, U. A. A internet e seus impactos na recuperação da informação. **Ciência da Informação***.* v. 26, nº 1, 1997.

TOUTANOVA, K.; KLEIN, D.; MANNING, C.; SINGER, Y. *Feature-Rich Part-of-Speech Tagging with a Ciclic Dependeny Network*. ***Proceedings of HLT-NAACL.*** 2003 p. 252-259.

TRASK, R.L. 2004. Dicionário de Linguagem e Linguística. **Contexto**, São Paulo, p. 364, 2004.

USC, 2008. **Universidade de Santiago de Compostela**. Disponível em: < http://gramatica.usc.es/~gamallo/parser_multilingua/index.htm >. Acesso em: 07 julho 2011.

USC. **Universidade de Santiago de Compostela**. Disponível em: < http://gramatica.usc.es/pln/tools/deppattern.html >. Acesso em: 09 julho 2011.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos.** 2004. Tese (Doutorado em Ciência da Computação), Universidade Federal do Rio Grande do Sul, 2004.