IURI DEOLINDO NOGUEIRA

DESENVOLVIMENTO DE UM SOFTWARE DE COMUNICAÇÃO ONLINE ENTRE DIFERENTES IDIOMAS COM TRATAMENTO DE EXPRESSÕES E LINGUAGENS ENCONTRADAS EM AMBIENTES DE CHATS

LAVRAS - MG
2012

IURI DEOLINDO NOGUEIRA

DESENVOLVIMENTO DE UM SOFTWARE DE COMUNICAÇÃO ONLINE ENTRE DIFERENTES IDIOMAS COM TRATAMENTO DE EXPRESSÕES E LINGUAGENS ENCONTRADAS EM AMBIENTES DE CHATS

Monografia apresentada ao Colegiado do Curso de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Orientador

Prof..Dr. Denilson Alves Pereira

LAVRAS
2012

IURI DEOLINDO NOGUEIRA

DESENVOLVIMENTO DE UM SOFTWARE DE COMUNICAÇÃO ONLINE ENTRE DIFERENTES IDIOMAS COM TRATAMENTO DE EXPRESSÕES E LINGUAGENS ENCONTRADAS EM AMBIENTES DE CHATS

Monografia apresentada ao Colegiado do Curso de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

APROVADA em 25 de Outubro de 2012

Profa. Dra. Marluce Rodrigues Pereira — Universidade Federal de Lavras

Prof. Dr. José Monserrat Neto — Universidade Federal de Lavras

Prof. Dr. Denilson Alves Pereira
Universidade Federal de Lavras
Orientador

LAVRAS
2012

Aos meus pais por insistirem, acreditarem e me apoiarem em todos meus sonhos e conquistas.

AGRADECIMENTOS

Primeiramente a Deus por ter me dado à chance de estar presente buscando o melhor, e principalmente por estar ao meu lado sempre.

Ao meu mentor, por toda proteção e força concedidos a mim.

Ao meu irmão por todos os momentos juntos e todas as felicidades que compartilhamos.

À minha namorada pela paciência, dedicação e força de vontade em me aturar e sempre buscar o melhor para o nosso relacionamento.

A todos os amigos e familiares por todas as risadas, diversões e ajudas durante todo esse período.

À Universidade Federal de Lavras e a todos os professores do Departamento de Ciência da Computação pelos ensinamentos transmitidos.

Ao Professor Dr. Denilson Alves Pereira pela orientação, paciência e ensinamentos que foram essenciais e de grande valia para a realização deste trabalho.

RESUMO

A popularização da comunicação digital tem implicado em um grande uso de ambientes de bate-papo *online* e redes sociais. Os usuários destes utilizam um padrão rápido de comunicação escrita, levando assim, ao surgimento de palavras e expressões que possuem várias alterações em níveis léxicos e semânticos. Tais alterações vêm interferindo na vida e no cotidiano desses usuários. Diante dessas circunstâncias, constatou-se a necessidade do desenvolvimento de um sistema de comunicação online (*chat*) que realize o tratamento de gírias e expressões tipicamente encontradas nesses ambientes, além do tratamento de expressões regionais e a tradução das mensagens dos usuários para outro idioma. Este trabalho visa facilitar e melhorar a comunicação dos usuários criando bases de dados com gírias, expressões e regionalismos, propondo assim um sistema capaz de identificar e tratar esses casos. Avaliou-se a qualidade desse sistema, bem como de sua base de dados, com as métricas Precisão, Revocação e F1, por meio de três experimentos utilizando 700 frases coletadas de *webchats* e redes sociais.

Palavras-chave: Chats *online*, Corretor de Palavras, Processamento de Linguagem Natural, Máquinas de Tradução.

ABSTRACT

The popularization of digital communication has been implicated in large environments using online chat and social networks. The users of these environments using a standard quick written communication, thus leading to the emergence of words and expressions that have multiple changes in lexical and semantic levels. These changes have interfered in the life and routine of these users. Given these circumstances, there was a need to develop a system of online communication (chat) to undertake the treatment of slang and expression typically encountered in these environments, and treatment of regional expressions and the translation of user messages into another language. This work aims to facilitate and improving the communication of users creating databases with slang, expressions and regionalisms, thus proposing a system able to identify and treat such cases. It was evaluated the quality of this system, as well as its database, with the metrics Precision, Recall and F1 through three experiments using 700 sentences collected from webchats and social networks.

Keywords: Online Chat, Natural Language Processing, Machine Translation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1. CONTEXTUALIZAÇÃO

A Internet é um conglomerado de redes de computadores interligados que oferece uma vasta gama de serviços e recursos para seus usuários. Dentre estes serviços, existem aqueles que possibilitam a comunicação entre dois ou mais usuários simultaneamente, que são os chamados chats (salas de bate-papo) e websites de redes sociais.

Devido ao caráter de interatividade e a popularização da comunicação digital, essas mídias tornaram-se altamente difundidas em poucos anos. Contudo, são formas de comunicação escrita muito rápidas e instantâneas. A necessidade de agilidade no diálogo levou ao uso de palavras e expressões que possuem várias alterações em níveis semânticos e léxicos [OTHERO, 2002]. A questão é que essa linguagem vem interferindo no trabalho cotidiano das pessoas que utilizam esses sistemas. Ou seja, internautas em geral tem perdido a noção de quando usá-la, e como consequência adotando uma linguagem informal para uso diário.

Segundo [GONÇALVES *et al*. 2006] a globalização tem implicado em um contato maior entre pessoas de diferentes nações e culturas. Essa comunicação, muitas vezes realizada por máquinas de tradução, não tem sido adequada. Termos abreviados, expressões regionais e gírias não são palavras que os tradutores conseguem traduzir com facilidade, e quando o fazem, muitas vezes não o fazem de forma correta.

Com base nesse escopo, constatou-se a necessidade de criação de uma ferramenta que interprete e trate palavras e expressões encontradas em ambientes de chats, realizando o processo adequado de correção da linguagem, com o objetivo de auxiliar a comunicação entre pessoas de diferentes idiomas

ou pessoas não acostumadas a utilizar esses termos.

1.2. OBJETIVOS DO TRABALHO

Este trabalho tem como objetivo principal o desenvolvimento de um sistema de chat *online* que possa conectar pessoas de diferentes nações, realizando a tradução das mensagens simultaneamente para seus respectivos idiomas, tratando expressões e gírias normalmente encontradas em ambientes de chats. Tem-se como objetivos específicos:

- Desenvolvimento de um chat *online* com opções para tradução de expressões entre os idiomas Português e Inglês;
- Criação de uma base de dados com expressões, gírias e linguagens encontradas em chats;
- Estudo de técnicas na área de processamento de linguagem natural para captura e tratamento adequado das mensagens do chat;
- Utilização de uma estrutura léxica para expansão de abreviações, de forma a auxiliar no processo de tradução da linguagem.
- Integração de uma API de tradução de linguagem adequada para o projeto.
- Avaliação da qualidade do sistema desenvolvido.

1.3. MOTIVAÇÃO DO TRABALHO

A primeira motivação fundamenta-se na influência que as mensagens vêm exercendo na vida cotidiana das pessoas. Segundo [OTHERO, 2002], a linguagem dos internautas está repleta de expressões novas, palavras

estrangeiras, abreviações, neologismos, entre outros. Como consequência dessa grande distorção da língua Portuguesa, acredita-se no surgimento de um subconjunto da norma padrão da mesma [FREITAG *et al*. 2006]. Esse novo padrão seria uma limitação do raciocínio dos usuários, já que os discursos utilizados caracterizam-se por serem curtos e abreviados.

A segunda motivação trata a respeito do uso de Sistemas de Recomendações e Mineração de Dados, que dependem exclusivamente de textos a serem analisados para seus respectivos objetivos. Para tanto, a falta de um texto redigido sem erros gramaticais é importante, o que nem sempre acontece, pois os dados podem não estar disponíveis em um formato simples de ser utilizado.

Diante dos motivos apresentados, o sistema proposto tem por objetivo ensinar a seus usuários o uso de expressões e palavras de maneira gramaticalmente correta, e as mesmas auxiliarão em melhores resultados para sistemas de recomendações e de mineração de dados.

1.4. ORGANIZAÇÃO DO TRABALHO

Este trabalho está estruturado da seguinte forma: O Capítulo 2 apresenta o Referencial Teórico, mostrando os principais conceitos utilizados para o desenvolvimento e construção do sistema, e trabalhos relacionados ao objetivo deste. No Capítulo 3, é mostrada a Metodologia de trabalho e os tipos de pesquisa aplicados. No capítulo 4, é descrito o processo de desenvolvimento do sistema, bem como sua estrutura. Seguindo para os Capítulos 5 e 6, têm-se, respectivamente, os resultados obtidos, a conclusão e trabalhos futuros.

2. REFENCIAL TEÓRICO

2.1. PROCESSAMENTO DE LINGUAGEM NATURAL

O Processamento de Linguagem Natural (PLN) é o campo da ciência que estuda a compreensão da informação por meio de um conjunto de regras e métodos formais com o objetivo de fornecer ao computador a capacidade de entender e compor textos. Essa compreensão de textos consiste no reconhecimento de seus contextos, realização de análises léxicas, sintáticas, semânticas, extração de informações, aprendizagem de conceitos, entre outros [SOARES, 2008].

Do ponto de vista da Ciência da Computação, a principal meta do PLN é construir sistemas computacionais que facilite a comunicação de forma efetiva, via linguagem natural, através dos conhecimentos linguísticos implementados em aplicações computacionais [VINHAES, 2005].

Além disso, as ferramentas de PLN também podem ter um complemento além de extrair informações, pois podem ser aplicadas para auxiliar a construir textos livres que podem ser oriundos de digitação ou de transformações como gravações de voz [BULEGON e MORO, 2010].

[SILVA *et al.* 2007] nos mostra através da Figura 1 que um sistema computacional interpreta uma sentença em linguagem natural através da análise de informações morfológicas (léxicas), sintáticas (regras gramaticais) e semânticas (significados), armazenadas em um dicionário. As delimitações das etapas acima ainda podem apresentar outras duas etapas: Integração de Discurso e Pragmática, que respectivamente, considera sentenças anteriores e posteriores a uma frase processada por um analisador semântico e reinterpreta a frase de forma que seu significado real seja definido.

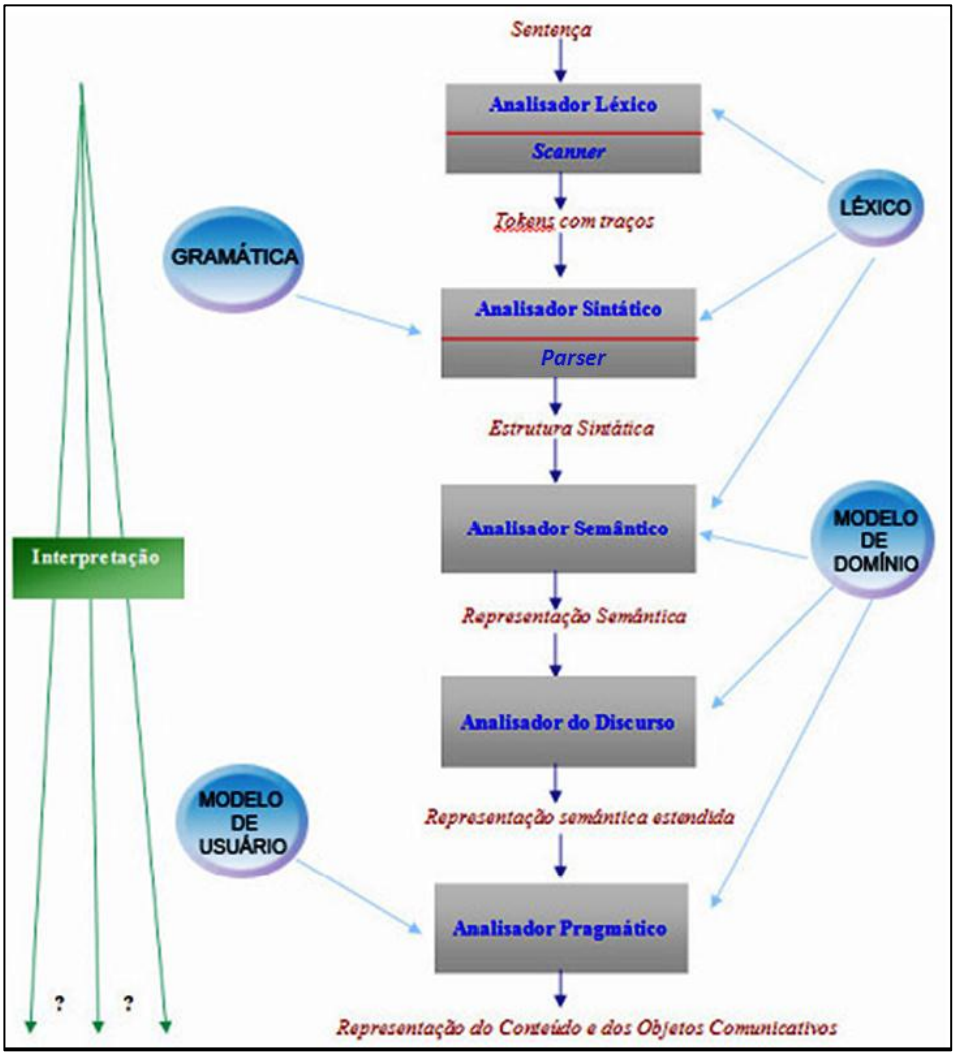

Figura 1: Sequência de Etapas do Processamento de Linguagem Natural em sistema computacional segundo [SILVA *et al*. 2007].

Antes que um texto comece a ser processado por um sistema de PLN, ele necessita passar por um pré-processamento, responsável pela preparação do mesmo, para que ele possa ser manipulado pelos algoritmos seguintes.

Para o presente trabalho, iremos utilizar a técnica de *tokenização* na fase de pré-processamento e um léxico na fase de Análise Morfológica ou Léxica, justificando-se, pois, o porquê de não adentrarmos em todas as fases de um sistema de PLN.

A técnica acima descrita irá nos fornecer os *tokens* das mensagens digitadas pelos usuários no sistema. Todas as correções e tratamentos das gírias e expressões serão trabalhados sobre esses *tokens*. Esse processo, o significado da palavra *token* e o léxico serão explicado adiante.

2.1.1. ANÁLISE MORFOLÓGICA OU LÉXICA

Segundo [OLIVEIRA, 2002], o analisador morfológico identifica palavras ou expressões isoladas em uma sentença, sendo este processo auxiliado por delimitadores (pontuação e espaços em branco). Cada uma dessas palavras ou expressões é um *token*. A estes *tokens* são dadas classificações quanto ao seu tipo de uso ou quanto sua categoria gramatical. O sistema ainda deve dispor de um dicionário (base de dados) onde serão armazenadas as palavras junto as suas informações e símbolos determinados pelas suas classificações.

Nesta fase ainda é possível a substituição da instância de uma palavra por outra do mesmo tipo, desde que se mantenha a sentença válida. As trocas só serão válidas se dentre os vários grupos de palavras com suas respectivas classificações, apenas palavras de grupos iguais podem ser substituídas. Dessa maneira, a análise morfológica trata de maneira congruente os grupos definidos e as palavras quanto suas formas e estruturas.

Para este trabalho, é imprescindível que esta fase seja executada no seu melhor desempenho, já que a mesma afetará toda a continuidade do projeto e

pode vir a convergir em vários erros. Assim, as técnicas que serão utilizadas são apresentadas a seguir.

2.1.2. *TOKENIZAÇÃO*

O processo de *Tokenização* é o primeiro passo na fase de pré-processamento. É responsável pelo seccionamento de um texto em unidades mínimas (*tokens*), mantendo a semântica original do texto. A sequência de *tokens*, por sua vez, é chamada de *tokenstream* [ARANHA, 2007].

A seleção da sequência de caracteres é auxiliada por fronteiras delimitadas por caracteres primitivos como espaço (“ ”), vírgula, ponto etc. Apesar que para [CARRILHO, 2007], os delimitadores podem assumir vários papéis, o que pode acarretar certos problemas. Como exemplo, o “ponto” (“.”) pode representar o fim de uma sentença, uma abreviação ou mesmo uma URL.

Como exemplo do processo, temos o seguinte exemplo: “O jogador de futebol, se aposentou.”. O resultado da técnica é uma sequência de palavras intercaladas por espaço e algumas vezes por símbolos delimitadores: [O] [jogador] [de] [futebol] [,] [se] [aposentou].

[KONDACHY, 2006] apresenta sua própria metodologia para a identificação de *tokens*, a qual é apresentada na Figura 2. Sua metodologia procura manter o alto padrão do nível semântico com a ajuda de dicionários de dados e regras de formação de palavras. Nota-se como ele propõe uma estrutura que engloba vários procedimentos para uma boa recuperação dos tokens.

A proposta de [KONDACHY, 2006] se torna uma importante referência diante do intuito deste projeto na correção de gírias e expressões encontradas em ambientes de chats, já que se dá a atenção necessária à abreviações e símbolos da Internet.

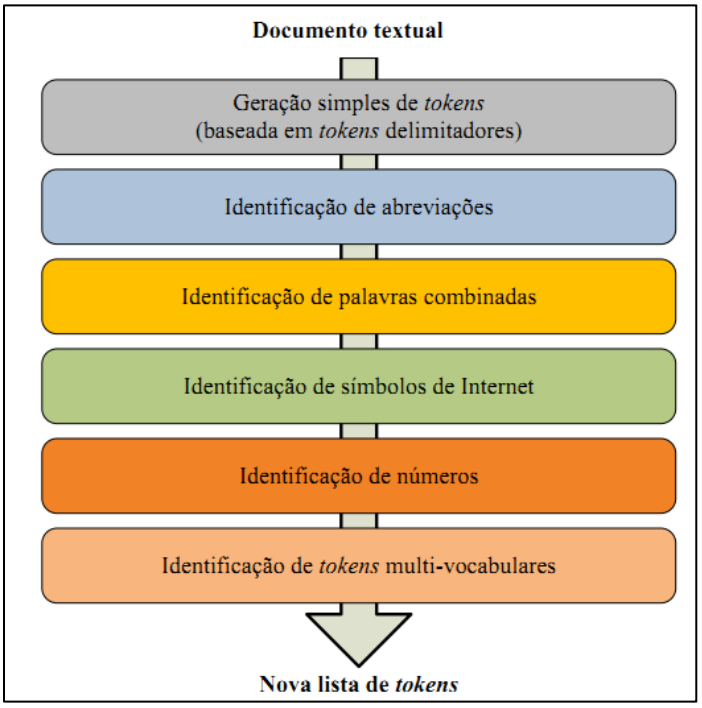

Figura 2: Metodologia de [KONDACHY, 2006] para a *tokenização*.

2.1.3. O LÉXICO

Segundo [NUNES, 1996], um léxico é um conjunto abrangente de palavras de um determinado idioma, onde cada palavra possui uma série de informações (atributos) sobre si, tornando-se possível a correção ortográfica e gramatical de textos desse idioma.

[ARANHA, 2007] ainda completa informando que um léxico pode ser representado por uma tabela, onde cada registro apresenta um lexema e seus atributos. Sobre esse formato de representação, a ideia é que cada registro da tabela apresente um lexema com um determinado significado.

O léxico será responsável por armazenar as gírias e expressões que serão tratadas durante a execução do sistema proposto, bem como seus devidos significados e ou correções.

2.2. MÁQUINAS DE TRADUÇÃO

As Máquinas de Tradução (MT) ou *Machine Translation* são programas de computador capazes de traduzir um texto de um idioma para outro. O seu estudo se encaixa na pesquisa e exploração de mecanismos nas áreas de PLN, Linguística Computacional, Inteligência Artificial (IA) e outros.

Segundo [HUTCHINS e SOMERS, 1992] o principal objetivo de uma MT não é apenas realizar a tradução dos textos, mas sim realizar uma tradução que seja possível manter a coerência e precisão. Apesar de um dos problemas aceitáveis é que grande parte das vezes não existe apenas uma boa tradução ou apenas um jeito de traduzir um texto, acrescentado do fato que diferentes pessoas podem possuir diferentes opiniões sobre uma boa tradução [PAPIPENI *et al*. 2002].

[VAUQUOIS, 1968] tentou clarificar as possíveis classificações de uma MT. A Figura 3 apresenta o Triângulo de Vauquois, que mostra que uma das possíveis classificações para uma MT é relativa à sua posição dentro do triângulo.

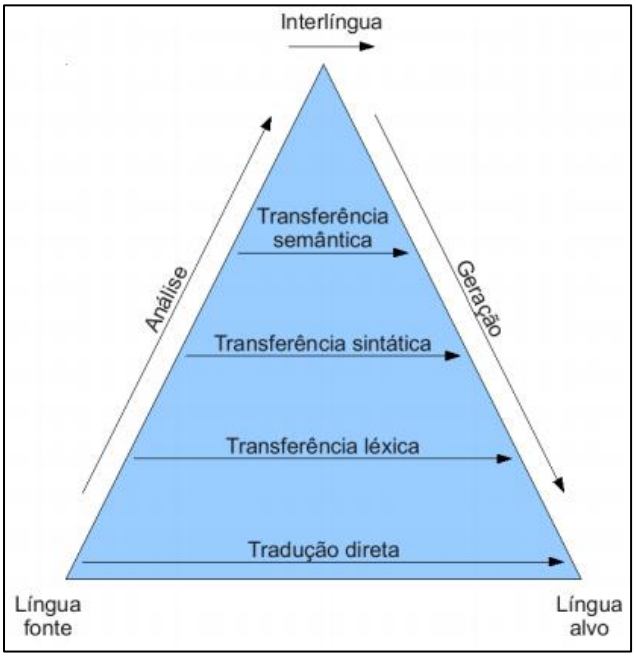

Figura 3: Triângulo de Vauquois para classificação de uma MT [VAUQUOIS, 1968].

2.2.1. ABORDAGENS DAS MT

2.2.1.1. BASEADA EM REGRAS

[SILVA, 2010] define Sistemas de MT baseados em regras como o princípio de que a tradução de um texto acontece a partir de um mapeamento sintático e um conhecimento preciso de uma linguagem alvo.

São modelos relacionados ao conhecimento de padrões e regras formais utilizando de recursos estruturados, como dicionários e ontologias, produzindo grandes resultados desde que sejam executados em conjuntos finitos. Suas abordagens são apresentadas a seguir de acordo com [BECK, 2009].

- **Tradução Direta:** A tradução é realizada palavra a palavra e em seguida acontece uma reordenação das palavras buscando organizar o

texto segundo a linguagem alvo. Utiliza de grandes dicionários e uma de suas vantagens é que ao não realizar, por exemplo, a análise semântica, o processo não fica sujeito a estes erros, contudo é difícil de detectar reordenamentos longos. A Figura 4 apresenta o esquema de uma tradução direta para duas linguagens.

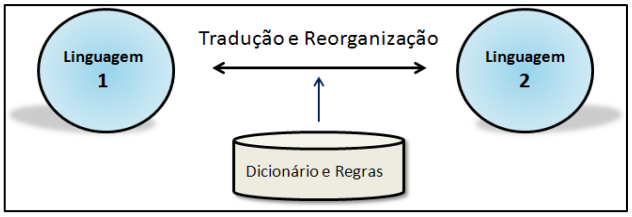

Figura 4: Tradução Direta para duas linguagens.

- **Tradução por Transferência:** O texto fonte passa por alguns processos de abstração e é transformado segundo a representação do grau que ele estiver. Esses graus podem ser de níveis morfológico à nível semântico [BECK, 2009]. Assim, torna-se simples a detecção de reordenação a longa distância e evita-se ambiguidades. A Figura 5 apresenta o esquema de uma tradução por transferência para duas linguagens.

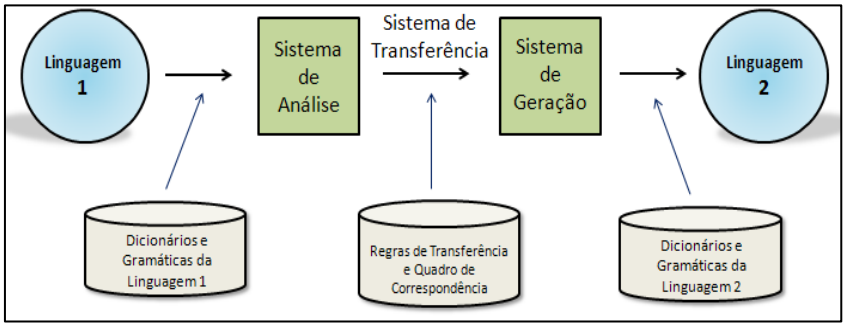

Figura 5: Tradução por Transferência entre duas linguagens. Existem dois dicionários e gramáticas, um para a linguagem um (1) e outro para a linguagem dois (2).

- **Tradução por Interlíngua:** Este modelo consiste em duas fases: na primeira fase, a frase da linguagem fonte é analisada e cria-se então uma representação dela para uma linguagem própria (interlíngua) que é independente das outras; na segunda fase a representação semântica é convertida para a linguagem de destino. A vantagem do sistema é que se cria apenas um módulo de análise e geração, contudo são difíceis de serem implementadas e de se colocar em prática devido a interlíngua. A Figura 6 apresenta o esquema de tradução por Interlíngua.

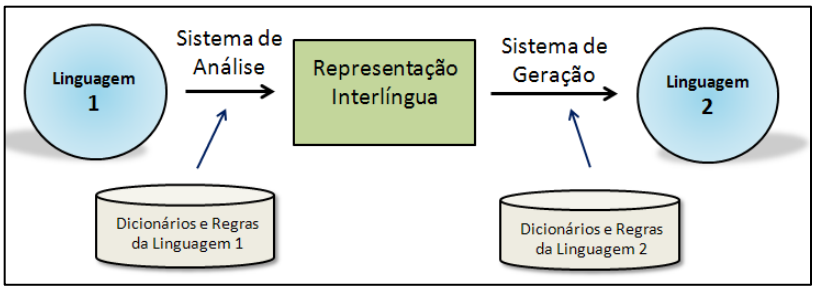

Figura 6: Tradução por Interlíngua entre duas linguagens. Encontra-se no meio da figura a linguagem própria, também chamada de representação interlíngua.

2.2.1.2. BASEADA EM DADOS

Em sistemas baseados em dados o princípio é que, para se realizar uma tradução, o tradutor possuirá dados prévios que utilizará para comparação, ou seja, vários exemplos de traduções são armazenados para posteriormente, serem usados como modelos para criação de novas traduções. Esses sistemas utilizam de técnicas de aprendizagem para selecionar o que é viável no momento de tradução.

Por possuir um alto nível de precisão nas traduções, tem sido mais utilizado que as outras abordagens [TRIPATHI e SARKHEL, 2010]. Suas abordagens são apresentadas a seguir.

- **Tradução por Estatística:** São caracterizadas pela utilização de métodos de aprendizagem de máquina, ou seja, aplica-se um algoritmo em uma grande quantidade de dados, textos e informações previamente traduzido, chamado de “corpus paralelo”. Assim, após aprender este corpus, o algoritmo é capaz de escolher boas traduções para frases inéditas [LOPEZ, 2008].

- **Tradução Baseada em Exemplos:** Este método é baseado em encontrar exemplos análogos na linguagem alvo. Ou seja, encontra exemplos da linguagem alvo que contribuirão para a tradução de acordo com sua similaridade com a linguagem fonte [DIETZEL, 2007]. Para verificar em números qual a melhor tradução, pode-se utilizar de comparação caractere a caractere. Destaca-se quando trabalha com pequenas sentenças, e principalmente com sentenças muito usadas, que podem ser traduzidas de maneira rápida.

2.2.2. A MT NESTE TRABALHO

Para o desenvolvimento de um ambiente de chat online conjunto ao sistema de tratamento de gírias e expressões é necessária a inserção de uma máquina de tradução capaz de realizar um tratamento adequado às diversas palavras encontradas.

Assim sendo, optou-se por utilizar uma máquina de tradução com estratégia em tradução por estatística, por ser a metodologia atual que apresenta

melhores resultados [TRIPATHI e SARKHEL, 2010]. Essa máquina de tradução escolhida foi o *Bing Translator*, conforme justificado a seguir.

2.3. API DE TRADUÇÃO

Uma *Application Programming Interface* (API), traduzido como "Interface de Programação de Aplicativos" é um conjunto de funções padronizadas que são oferecidas aos programas que as solicitam.

Em [KIT e WONG, 2008], apresenta-se estudos recentes de alguns tradutores *online* gratuitos. Entre os tradutores, encontram-se Babelfish (babelfish.yahoo.com), Google Translate (www.translate.google.com), Systran (http://www.systran.co.uk/), e outros. Todos os tradutores foram avaliados em duas pontuações diferentes, com dois tipos de dados. Para os idiomas Inglês e Espanhol, o Google Translate e o Babelfish foram os únicos que apresentaram melhores resultados.

A API do Google Translate utiliza máquinas de tradução baseadas em estatística para suas traduções. No momento que necessitar traduzir um texto, o Google varre uma enorme quantidade de documentos procurando por padrões parecidos em documentos que já tenham sido traduzidos e revisados por tradutores humanos. De posse de todos os documentos, ele utiliza da abordagem estatística para selecionar qual a melhor tradução (http://translate.google.com/about/intl/en_ALL/).

Uma API de tradução de linguagem é utilizada na implementação deste trabalho. Para a escolha desta API baseou-se na disponibilidade da API para uso gratuito. Como o Google exige um taxa para o uso de sua API, foi escolhida a

API do Yahoo para ser usada neste trabalho, que é gratuita. A seguir apresenta-se o seu funcionamento.

Recentemente, o Yahoo adotou o *Bing Translator* (http://www.bing.com/translator), substituindo a API Babelfish, como seu tradutor oficial, visando melhorar seus serviços. Essa tecnologia de tradução foi desenvolvida pela *Microsoft Research*. O *Bing Translator* foi construído com o intuito de ser uma máquina de tradução que aprenda automaticamente os mapeamentos da tradução de dois corpora de linguagens diferentes. Utilizando desta estratégia, seus criadores adotaram uma abordagem orientada a dados, ou seja, treina-se o algoritmo através de textos traduzidos e corrigidos por humanos. Esse tipo de estratégia coincidiu com tradução automática estatística, que juntas, deram forma ao tradutor [MENEZES e QUIRCK, 2005].

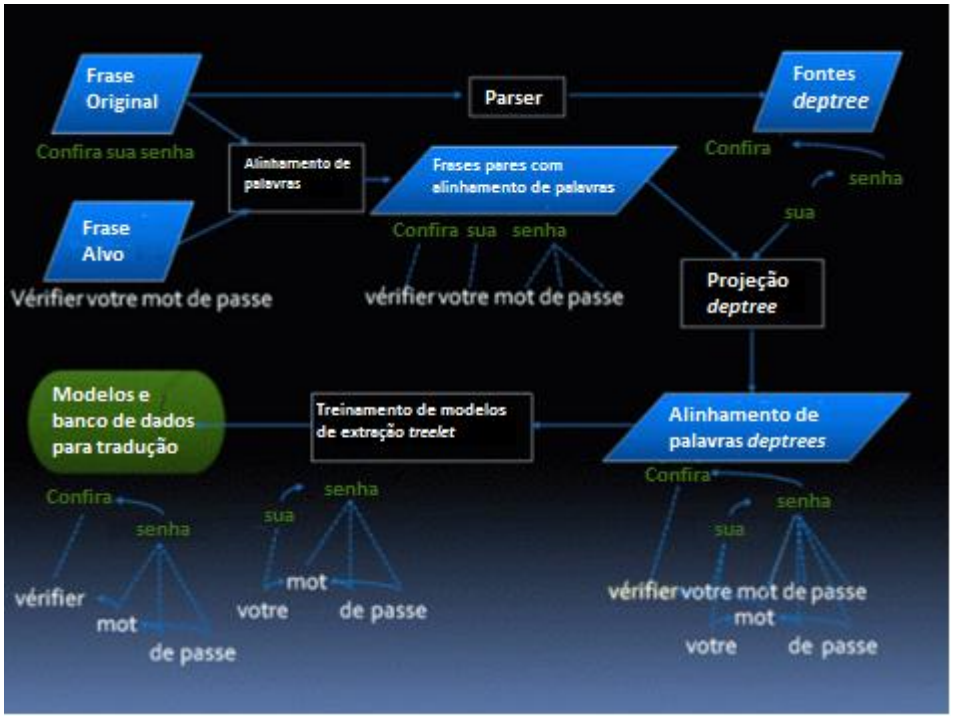

Figura 7: Funcionamento do sistema da API do *Bing Translator*.

A Figura 7 mostra o funcionamento do sistema da MT. A entrada do sistema recebe a sentença original e a sentença já corrigida por tradutores humanos. Realiza-se todo seu processamento gerando por fim modelos pré-estabelecidos de traduções, que em momento de execução, através de métodos estatísticos, apenas um será escolhido como tradução mais adequada.

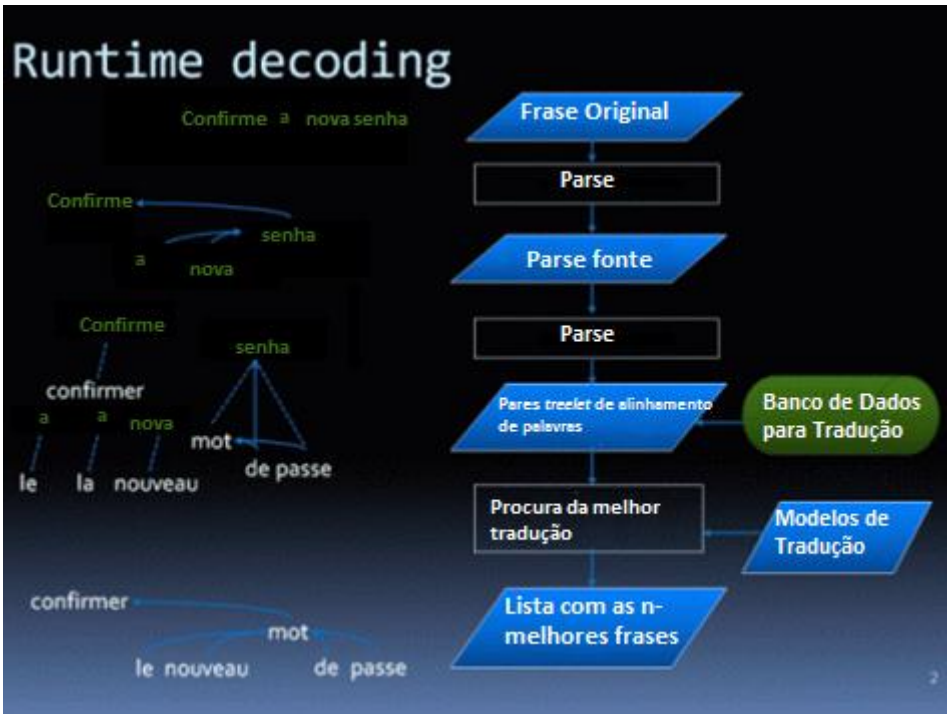

Figura 8: Exemplo de tradução da frase "Confirme a nova senha" do idioma Português para o idioma Francês de acordo com o *Bing Translator.*

A Figura 8 apresenta o sistema em tempo de execução para o exemplo de tradução de uma frase em Inglês para Francês de acordo com o tradutor *Bing Translator*. Através da figura, nota-se que após ser aplicado o *parser* na sentença original, aplica-se um sistema de alinhamento das palavras e por fim,

usa-se de métodos estatísticos em modelos pré-definidos para encontrar a melhor tradução.

2.4. TRABALHOS RELACIONADOS

[ANACLETO *et al*. 2010] desenvolveram um trabalho para lidar com usuários de diferentes culturas, buscando auxiliá-los no processo de criação de mensagens para outros idiomas através de um chat, que serve como uma extensão de uma ferramenta de tradução automática.

Eles trabalharam com uma base de dados de conhecimento do senso comum, ou seja, eles coletam as sugestões dos usuários de acordo com suas respostas para sentenças semiestruturadas. Sentenças estas, compostas por três partes: estática, dinâmica e resposta do usuário.

Para determinar o valor da resposta do usuário, eles utilizam o conceito de relacionamento de Misnky [MINSKY, 1986]. Segundo o autor, o conhecimento humano pode ser mapeado através de 20 relacionamentos, como por exemplo: “definido como”, “é um”, “parte de”, “feito de”, entre outros. Esses e outros relacionamentos são as partes estáticas das sentenças disponíveis para os usuários. A parte dinâmica consiste em palavras automaticamente preenchidas pelo computador.

Assim sendo, eles formam sentença, por exemplo: “Futebol é um ____”. Onde “futebol” é a parte dinâmica, “é um” é um dos relacionamentos de Minsky e “___” é a parte de resposta do usuário.

A partir do momento que o usuário preenche a sentença, a mesma é inserida na base de dados e é processada através de mecanismos de PLN, que verificam qual a veracidade da sentença. Esse esquema pode ser notado na

Figura 9, onde é possível visualizar o uso de duas aplicações computacionais do projeto de [HAVASI *et al.* 2007]: *"MCS Knowledge Database"* e "*ConceptNets"*. Este projeto, nomeado "Open Mind Common Sense" tem por objetivo a coleta de conhecimento através da opinião das pessoas sobre vários assuntos. E esse conhecimento, posteriormente, utilizado no desenvolvimento de aplicações computacionais que sejam úteis as pessoas.

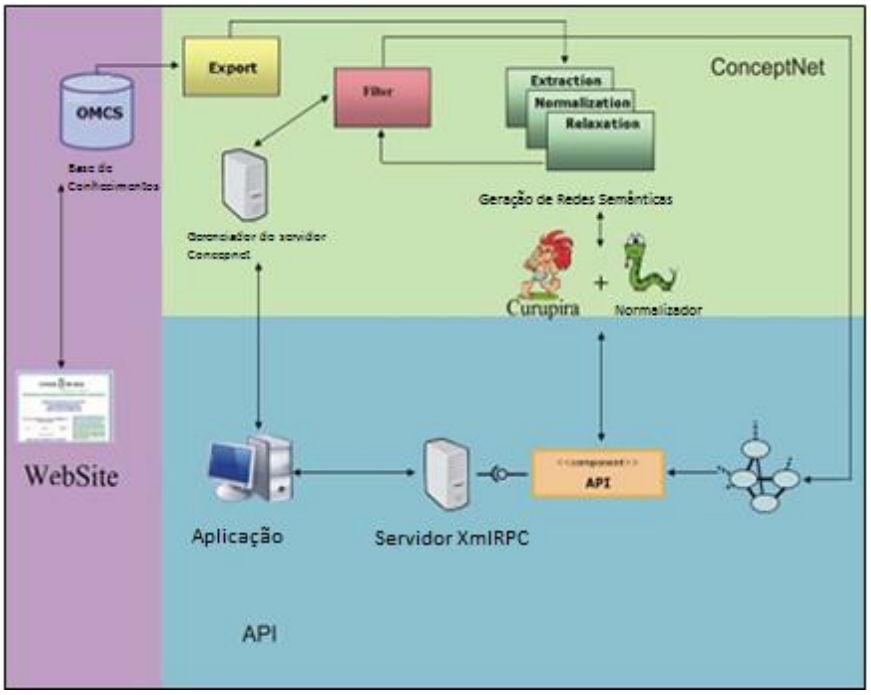

Figura 9: Esquema do processamento de sentenças de acordo com o projeto de [ANACLETO *et al*. 2010].

O *OMCS Knowledge Database* nada mais é que o banco de dados dos conhecimentos adquiridos das pessoas que acessam seu sistema e deixam suas opiniões. A coleta de informações parte do princípio do conceito de relacionamento de Misnky [MINSKY, 1986], já explicado.

O *ConceptNets* consiste em uma rede semântica com base nas informações do banco de dados. É expresso através de grafos direcionados que possuem nós representando os conceitos, e arestas com seus respectivos relacionamentos. Esses conceitos podem ser substantivos, verbos, adjetivos, entre outros.

Ao contrário do trabalho de [ANACLETO *et al*. 2010], este trabalho pretende dar em enfoque maior em relação às gírias, diferindo deles que trabalham mais com a parte de expressões e reconhecimento de categorias das palavras.

Já no trabalho de [CASELI *et al.* 2010], os autores apresentam ideias sobre uma nova abordagem de uma MT. Busca-se verificar qual o efeito de se gerar traduções automáticas usando MT estatística aplicando a base de dados de conhecimentos em diferentes etapas do processo de tradução.

Para realizar seu trabalho, os autores utilizaram de duas redes semânticas derivadas do *ConceptNets* com informações dos idiomas inglês e português do Brasil. Através destas redes os autores pretendem aplica-las em três momentos: Antes da tradução com a intenção de enriquecer o texto com a base de conhecimentos auxiliando assim a máquina de tradução; durante a tradução automática onde a base de conhecimento é utilizada como uma nova funcionalidade da máquina de tradução e; depois da tradução automática utilizando a base de conhecimento para expor melhor o significado de certas palavras que não fossem bem traduzidas.

O trabalho não apresenta os resultados, pois o propósito do mesmo, no momento, era apenas divulgar a ideia e como essa nova abordagem será testada.

Em outro trabalho, o de [SUGIYAMA *et al.* 2011], os autores propõe um chat que ajude seus usuários a produzir mensagens no idioma inglês. Como proposta diferencial, eles incluem recursos que consideram a cultura de seus usuários, também conhecida como "Cultura Sensível ao Computador", buscando assim proporcionar uma interação mais natural para seus usuários.

O chat foi desenvolvido sobre uma abordagem centrada no usuário, utilizando da base de conhecimentos *OMCS Knowledge Database in Brazil* para seu banco de dados e também a API do Google Translate para as traduções. Tanto a base de dados cultural (*Cultural Translation*) como a API poderiam sugerir traduções, ficando a critério dos usuários usá-las ou não.

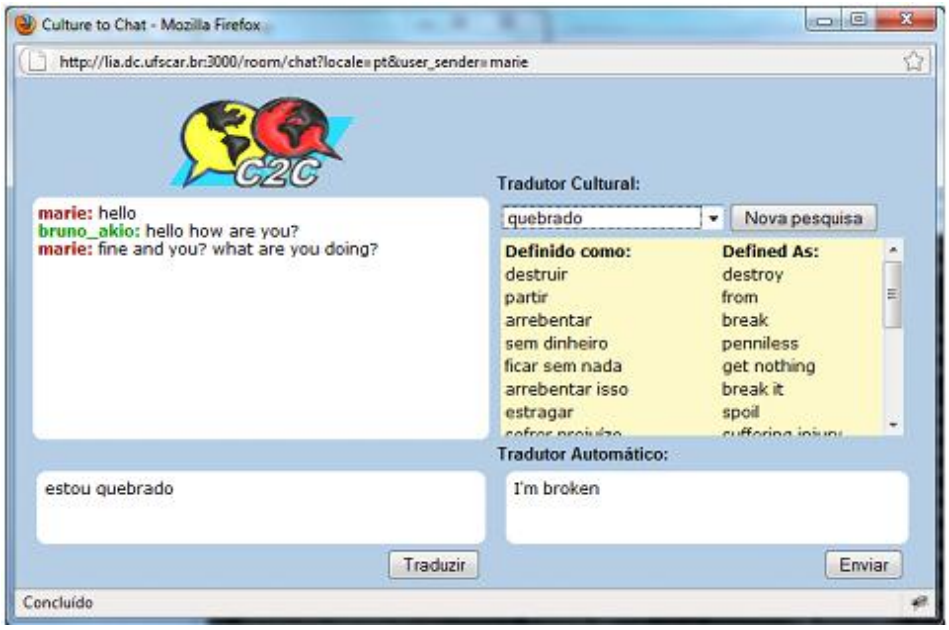

Figura 10: Interface do chat chamado "Culture-to-Chat" (C2C) proposto por [SUGIYAMA *et al.* 2010].

A Figura 10 apresenta a interface do chat que foi desenvolvido em “Ruby and Rails”. O chat apresenta na janela superior esquerda a conversa com o outro usuário; na janela inferior esquerda a mensagem que o usuário deseja enviar; na janela superior direita exemplos da base de dados cultural para traduções específicas; e na janela inferior direita a mensagem final que será enviada.

Como forma de avaliar o projeto, os autores realizaram um caso de estudo envolvendo usuários brasileiros e canadenses. Para os usuários brasileiros foram entregues questionários para identificar o nível de inglês que cada um apresentava e para os canadenses, questões de experiência em conversas com estrangeiros Brasileiros.

Os resultados mostraram um número muito pequeno do uso da tradução cultural em relação à quantidade de traduções realizadas pela API e pelo número total de mensagens enviadas. Vale ressaltar que a tradução cultural foi considerada a quantidade de vezes que os usuários utilizaram a busca direta nos exemplos da janela superior à direita. Outro método de avaliação importante foi tentar averiguar a evolução dos usuários de acordo com o chat. Verificou-se que para os usuários com baixo conhecimento no idioma inglês o chat foi considerado rápido, enquanto que para usuários mais avançados no idioma o chat se apresentou mais devagar.

Ao final dos testes, conclui-se que a base de conhecimentos cultural pode não estar tão apta a trabalhar com os coloquialismos, apesar de os usuários terem aprovado as traduções rápidas realizadas pela API.

Diferentemente do trabalho de [SUGIYAMA *et al.* 2011], este trabalho procura encontrar expressões mais utilizadas pelos usuários de chats e trabalhar em cima disso. Gírias e regionalismos são os enfoques deste buscando aperfeiçoar a interação entre pessoas de diferentes nacionalidades.

3. METODOLOGIA

A metodologia de pesquisa visa definir o que foi pesquisado neste trabalho. A pesquisa bibliográfica serve como base para a aquisição de conhecimento acerca dos temas envolvidos no projeto. Basicamente envolve consultas a livros de referência, teses científicas e artigos nas áreas de Recuperação da Informação, Processamento de Linguagem Natural e Máquinas de Tradução.

O estudo do problema foi o primeiro passo para a realização da pesquisa, seguido da aquisição de referenciais bibliográficos para a familiarização, embasamento científico do trabalho e identificação de problemas que são relacionados ao seu escopo.

Seguiu-se ao desenvolvimento e implementação do sistema de chat *online* integrado de módulos de tratamento de gírias e expressões e da API de tradução de idiomas.

Antes da realização dos testes, preencheu-se a base de dados com gírias e expressões previamente pesquisadas e coletadas. Definiu-se pois, quais seriam e como seriam executados os testes. Por fim, analisou-se os resultados concluindo-se este trabalho.

3.1. CLASSIFICAÇÃO DA PESQUISA

[JUNG, 2004] define que um trabalho científico pode ser avaliado segundo sua natureza, quanto aos seus objetivos e aos seus procedimentos. Essa avaliação serve para definir um escopo para a área de atuação dos trabalhos. Os tipos de pesquisa deste trabalho são apresentados a seguir.

Este trabalho caracteriza-se por ser uma pesquisa de natureza aplicada ou tecnológica, devido ao fato de utilizar-se de conhecimentos adquiridos na literatura para uma aplicação prática, implementando um sistema online com tratamento da informação vinculada aos ambientes de chats.

Quanto aos objetivos, a pesquisa pode ser classificada de caráter exploratório, pois buscou-se constatar algum problema em um fenômeno específico e familiarizar-se com o mesmo. Partindo deste ponto, tem-se a finalidade de encontrar e aprimorar ideias que trouxessem a resolução do problema de maneira que houvesse uma contribuição de alto padrão.

Por fim, quanto aos procedimentos, este trabalho caracteriza-se por ser uma pesquisa experimental, pois houve a implementação de um protótipo de software e realização de experimentos para a avaliação do sistema dentro de seu contexto.

3.2. PROCEDIMENTOS METODOLÓGICOS

A metodologia deste trabalho foi estruturada em seis etapas, quais sejam: estudo e consideração do problema, aquisição de referências bibliográficas para familiarização, desenvolvimento do chat *online*, desenvolvimento dos módulos separadamente, coleta de dados para averiguação do protótipo, aplicação e análise dos experimentos. Como módulos, considera-se as três etapas para o tratamento da mensagem. Estas etapas em ordem de execução são: tratamento de gírias; tratamento de expressões e; integração de uma API de tradução de idiomas.

Para os testes, definiu-se o uso dos conceitos de “Precisão e Revocação” nos resultados encontrados após o processamento do sistema em frases

coletadas de ambientes de *chats* e redes sociais. Analisou-se a qualidade da tradução averiguando sua precisão e também realizou-se uma análise de uso da base de dados.

3.3. TECNOLOGIAS ENVOLVIDAS

A principal ferramenta utilizada para o desenvolvimento do projeto foi o NetBeans Enterprise Edition (Java EE) IDE para desenvolvedores Web em sua versão 7.1.1. O NetBeans foi utilizado na criação dos arquivos de classes Java e JavaServer Pages.

O uso da programação em Java para desenvolvimento do sistema deu-se pelo motivo da plataforma Java proporcionar a facilidade de programação orientada a objetos, integração de sistemas com outras tecnologias, bem como a variedade na quantidade de bibliotecas oferecidas [BUYYA *et al*. 2009].

Utilizou-se o Apache Tomcat como servidor das aplicações Web, na sua versão 7.0.22. O Apache Tomcat é um servidor *open source*, e foi escolhido devido ao seu funcionamento independente do sistema operacional, facilidade de uso e por operar plataforma Java.

Foi necessário acrescentar ao projeto uma API de tradução. Essa API como foi relatada no Referencial Teórico, é a API do *Bing Translator* do Yahoo, devido principalmente a sua utilização gratuita.

O sistema de chat desenvolvido está disponível em um servidor Linux, fisicamente localizado no Departamento de Ciência da Computação da Universidade Federal de Lavras, e que pode ser acessado de qualquer computador conectado a Internet no endereço http://env-8305678.j.rsnx.ru/.

4. DESENVOLVIMENTO

Primeiramente, será dada uma visão geral do que foi feito e posteriormente, a apresentação detalhada da construção e das funções que o protótipo em questão possui.

4.1. VISÃO GERAL DO SISTEMA

Implementou-se um protótipo de um chat *online* para tratamento de gírias, expressões, regionalismos e realização da tradução das mensagens dos usuários do idioma português para o inglês.

O sistema exige do usuário um apelido e seu idioma de trabalho. Esse idioma será utilizado posteriormente para tradução, se necessário. O usuário entrando com esses dados possuirá acesso ao chat. No chat o usuário poderá enviar mensagens a cada outro usuário individualmente ou mesmo a todos em conjunto.

Cada mensagem que o usuário envie, passará por um tratamento, corrigindo e expandindo as gírias e expressões. Caso seja identificado que um usuário envia uma mensagem para outro usuário de outro idioma, a mesma será traduzida pelo módulo de tratamento encarregado pela API de tradução de idiomas.

O sistema ainda conta com uma opção para o usuário inserir uma gíria ou expressão. Contudo, evitando que falsas gírias sejam inseridas, o sistema analisa se a mesma expressão já foi ou não sugerida anteriormente.

4.2. INTERFACE DO SISTEMA

A Figura 11 apresenta a tela inicial do sistema. Essa tela contém dois campos. No primeiro campo, o usuário entra com o apelido que ele utilizará no chat, e no segundo campo, ele identificará o seu idioma.

Figura 11: Interface da página principal do chat.

Já a Figura 12 é a tela do chat que aparece logo após o usuário ter informado ao sistema seu apelido e idioma. Nela o usuário poderá visualizar as informações que acontecem no chat, suas mensagens enviadas e recebidas. Também poderá escolher outro usuário *online* para enviar uma mensagem e o mais importante, poderá ver como sua mensagem será enviada após o devido tratamento das gírias, expressões e tradução, caso seja necessário.

Toda vez que o usuário desejar enviar uma mensagem, ele pode enviar a mensagem diretamente, ou pode clicar no botão de pré-processamento para que a mensagem seja devidamente tratada. Após esse tratamento, o usuário receberá uma sugestão de mensagem, onde o mesmo poderá ou não modificá-la de acordo com sua opinião, e por fim enviá-la.

Figura 12: Interface da página de conversação dos usuários do chat.

A Figura 13 demonstra o procedimento de tratamento da mensagem de um usuário. A primeira etapa consiste no tratamento de gírias. Essas gírias foram coletadas em ambientes de chats. A segunda etapa é responsável pelo tratamento de expressões e regionalismos, como por exemplo: *feijoada* e *caipirinha*. Finalmente, tem-se a terceira e última etapa realizando a tradução da mensagem para o idioma do usuário alvo.

Ressalta-se que caso um usuário queira enviar uma mensagem para outro usuário do mesmo idioma, ocorrerá apenas a primeira etapa do processo de tratamento.

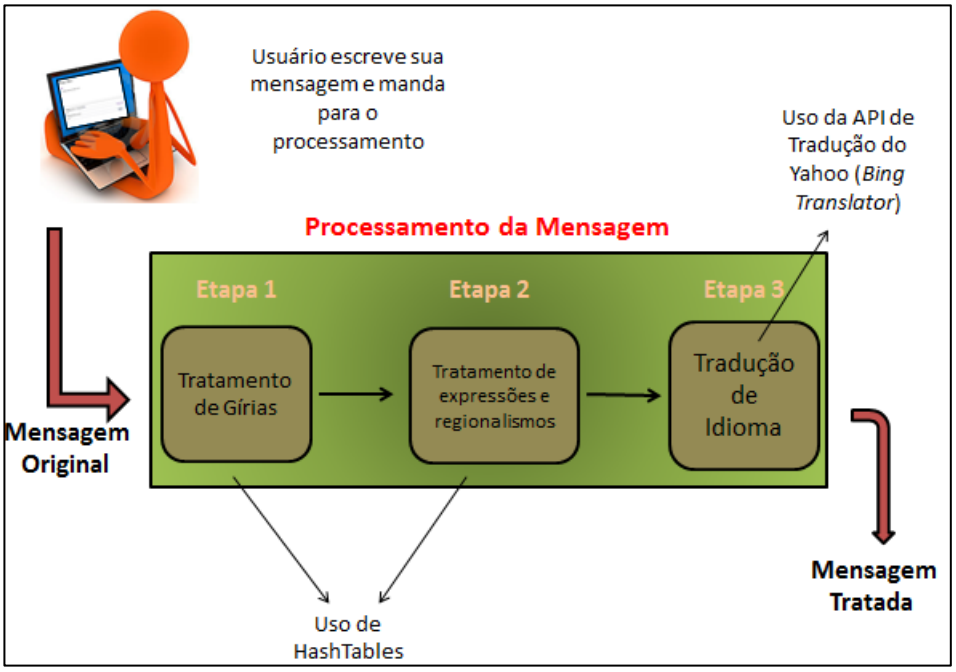

Figura 13: Etapas do procedimento de tratamento das mensagens. Na primeira etapa ocorre um tratamento das gírias. Na segunda etapa ocorre um tratamento de expressões e regionalismos. A última etapa consiste na tradução da mensagem para uma linguagem alvo.

4.3. O CHAT *ONLINE*

O chat foi desenvolvido sobre a plataforma J2EE (Java Enterprise Edition), fazendo uso de Servlet's e JSP. A arquitetura implementada segue o padrão de projetos MVC (Model View Controller), em que a camada de Visão foi construída utilizando de HTML e JSP. A camada de Controle foi implementada utilizando Servlets e a camada de Modelo são utilizadas classes Java.

A camada de Controle possui um único Servlet que recebe todas as requisições da aplicação, e de acordo com a regra de negócios recebida, invoca

o procedimento necessário gerando uma saída que é retornada ao Servlet, e este, ao cliente que realizou a requisição.

Para a invocação de uma dada regra de negócios foi criada uma interface chamada “BusinessLogic”, que possui um único método, para que seja padronizado o comportamento das classes que implementam as regras de negócios.

A execução de uma regra de negócios se dá como segue: ao receber uma requisição, o Servlet Controller lê a URL requisitada e extrai dela o nome da classe que deve executar a regra de negócio solicitada. Assim, por meio do padrão Reflection, é obtida uma instância dessa classe que é abstraída para a interface BusinessLogic, e então é executada.

Para o chat foram desenvolvidas as seguintes regras de negócios: Login e logout de usuários; Envio de mensagens; Janela de mensagens; Controle de usuários ativos; Recuperação das informações do formulário e Tratamento da Mensagem.

Na camada de Visão, além de HTML e CSS, também é utilizado o Framework jQuery para se trabalhar com Javascript e requisições assíncronas. A página do chat é composta por dois Frames, um onde as mensagens são mostradas e outro onde é apresentado um formulário para o envio de uma mensagem a outro usuário, com a opção de realização do Pré-processamento da mensagem. Uma importante característica do Frame onde as mensagens são exibidas é que a conexão HTTP deste Frame com a aplicação fica aberta e só se encerra quando o usuário sai da página do chat.

A camada de Modelo constitui o acesso e persistência dos dados de configuração da aplicação. O carregamento das configurações é feito no instante

que ocorre o primeiro acesso a aplicação, e são mantidas em uma única instância de um objeto. Para alcançar esse objetivo foi implementado o padrão de projetos “Singleton” dentro da classe de persistência das configurações da aplicação.

4.4. MÓDULOS DE TRATAMENTO DE GÍRIAS E EXPRESSÕES

Para o tratamento das gírias e expressões populares os módulos utilizam-se de duas estruturas *Hash*. A primeira responsável por armazenar as gírias, e a segunda por armazenar as expressões e regionalismos. A decisão do uso de tabelas *Hash* para implementação se deu por ser uma estrutura rápida para pesquisa.

Quando uma gíria ou expressão possui um relacionamento 1:1 com seu significado, o sistema simplesmente expande-a, corrigindo-a. Contudo, existem alguns casos que uma gíria possui relacionamento 1:n, onde n > 1, para com seus significados. Por exemplo, a gíria “c” que pode significar tanto “você”, quanto “se”, quanto “com”. Nestes casos, o sistema transfere a responsabilidade de decisão ao usuário apresentando-o todas as opções, fazendo-o escolher qual das expansões é a mais adequada.

Estes módulos trabalham com uma busca em níveis pelas gírias e expressõe, isto é, no nível 1, o método de *tokenização* recupera *tokens* palavra à palavra, no nível 2, os *tokens* são formados de duas em duas palavras, no nível 3 de três em três palavras e assim por diante, usando um esquema de janela deslizante sobre o texto.

A busca começa no nível 4, ou se houver um número menor de palavras na frase, no nível da quantidade de palavras. Cada um desses *tokens* é analisado e tratado, caso ele seja uma gíria ou expressão. O valor 4, determinado pelo

autor, foi escolhido por não haver nenhuma expressão ou gíria com quantidade superior a 4 palavras.

4.5. TRATAMENTO PARA ADIÇÃO DE NOVAS GÍRIAS E EXPRESSÕES

Existe no chat a opção do usuário poder adicionar uma nova gíria ou expressão à base de dados. Ao adicionar uma gíria o usuário entra com o nome da gíria e seu significado, por exemplo: nome = “vc” e significado = “você”. Contudo, para evitar falsos casos, o sistema realiza os três procedimentos descritos a seguir, antes de confirmar que aquela gíria ou expressão existe e significa aquilo que o usuário sugeriu.

O sistema utiliza-se de uma tabela específica para essas sugestões dos usuários. Caso as sugestões para uma mesma gíria aconteçam três vezes, por usuários diferentes, e sejam idênticas, tanto no nome quanto no significado, o sistema aceita a mesma, colocando-a na sua base de dados. A questão de se aceitar uma gíria ou expressão apenas quando se tem a frequência de três vezes, foi determinado pelo autor, que considera o três, um valor mínimo para evitar que usuários façam sugestões inválidas ou falsas.

O segundo procedimento que o sistema verifica é quando existem apenas divergências nos significados, mas possuindo o mesmo nome. São casos em que, por exemplo, as sugestões de dois usuários diferem por uma pequena palavra, ou mesmo letras. Neste caso, implementou-se o método da distância de *Levenshtein* [YUJIAN e LIU, 2007], que verifica quão uma palavra é diferente da outra, de acordo com a quantidade de mudanças de letras que uma palavra necessita para ser igual a outra. Essas mudanças podem ser adição, subtração ou troca de letras. O valor da mudança corresponde à porcentagem de letras

modificadas em relaçãoao número total de letras. Um exemplo é o caso da gíria “pd” apresentado na Figura 14.

Por fim, o último procedimento evitará que o mesmo usuário faça três sugestões iguais, verificando na tabela se é o mesmo usuário que realizou as demais sugestões.

Assim, quando três usuários sugerem uma mesma gíria ou expressão, o sistema irá analisar se entre essa nova gíria ou expressão e as já existentes existe uma porcentagem mínima de razão. Essa porcentagem mínima será um valor de no máximo 15%. Valor este testado em palavras com poucas letras, resultando neste valor mínimo para que duas palavras possam ter significados diferentes. Exemplifica-se uma situação desta porcentagem na Figura 14.

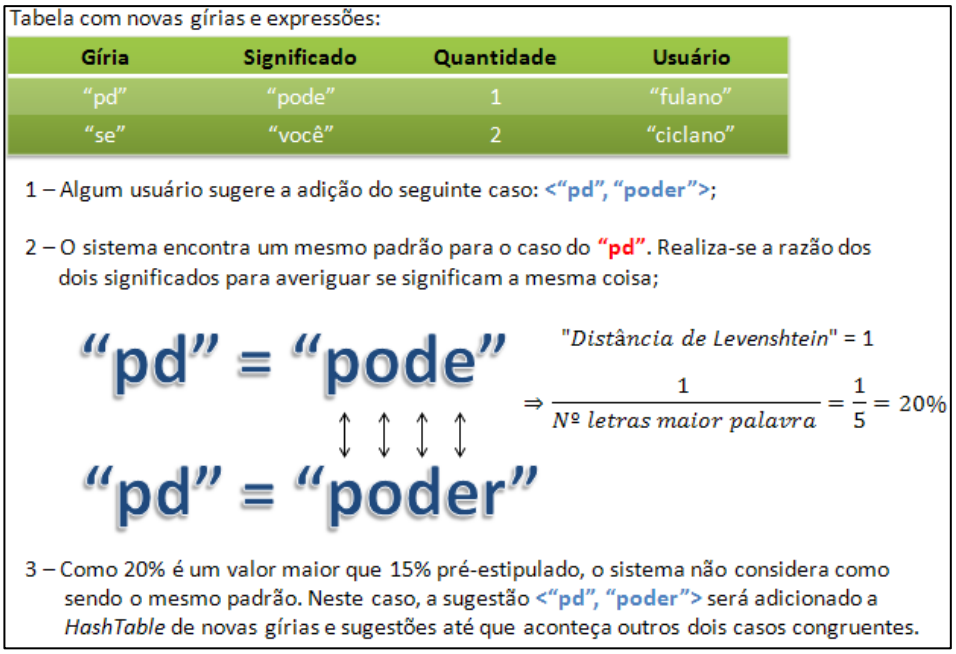

Figura 14: Demonstração de como o sistema irá reagir caso algum usuário sugira a adição de uma nova gíria ou expressão.

Para averiguar este valor, o módulo de adição de gírias ou expressão trabalha com a medida de “Distância de Edição” ou “Medida de Levenshtein” [ATALLAH, 1999]. O algoritmo dessa medida consiste em calcular o menor número de inserções, exclusões e substituições requeridas para mudar uma palavra inicial em outra. Depois de calculado, esse valor é dividido pela quantidade de letras da maior das duas palavras. Caso esse valor seja maior que 15%, ela não é considera igual.

No caso da Figura 14, os significados das duas palavras diferem e assim sendo, o conjunto <“pd”, “poder”> seria adicionado a tabela de novas gírias e expressões.

Ressalta-se que antes de verificar a sugestão na tabela de novas gírias ou expressões, o sistema busca nas tabelas *Hashs* já existentes. Caso já exista um caso idêntico, a sugestão do usuário é descartada. Caso apenas o significado seja diferente, o sistema realiza a análise para verificar se a razão é mínima. Se for, descarta a sugestão, se não for, irá buscar na tabela de novas sugestões.

A Figura 15 apresenta a interface específica para adição de gírias. O usuário apenas precisa preencher qual o nome e o significado da gíria.

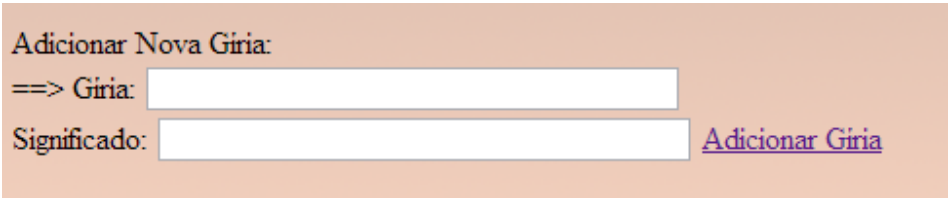

Figura 15: Interface para adição de novas gírias.

5. TESTES E RESULTADOS

Este capítulo apresenta os testes e seus respectivos resultados obtidos dos experimentos para se avaliar o sistema desenvolvido. Inicialmente é apresentado como realizou-se a coleta dos dados para a base de testes, seguido das métricas que foram utilizadas para avaliar o sistema. Depois, mostra-se a execução de cada teste, seus resultados e comentários de avaliação. Por fim, apresenta-se uma discussão geral sobre todos os resultados em conjunto.

Inicialmente, preencheu-se a base de dados, entre os dias 10 e 20 de abril de 2012, coletando-se gírias e expressões dos mesmos locais de onde foram coletadas as frases para teste. As gírias e expressões foram coletadas de frases aleatórias de conversas dos usuários dos chats, em diferentes horários e diversas salas, visando alcançar uma maior diversidade, sem que houvesse tendência para algum assunto específico. Foram coletadas 570 gírias e 368 expressões.

Junto à coleta, investigou-se através dos chats e sites específicos qual o real significado de cada uma das gírias e expressões coletadas. As gírias e expressões coletadas, bem como seus respectivos significados encontram-se disponíveis na página inicial do sistema.

Como forma de se verificar a qualidade do sistema proposto e a sua real contribuição, realizou-se testes utilizando-se dois conjuntos de frases encontradas em ambientes de chats.

O primeiro conjunto consiste de um total de 350 frases coletadas em salas de bate-papo *online*. Essas frases foram coletadas de frequentados chats brasileiros, que são: Bate-papo da UOL (http://batepapo.uol.com.br/), BOL (http://bpbol.uol.com.br/), IG (http://batepapo.ig.com.br/) e Terra

(http://chat.terra.com.br/), entre os dias 23 e 30 de abril de 2012. A este conjunto de frases, nomeia-se grupo “um” (1).

O segundo conjunto consiste de 350 frases coletadas no mesmo período de coleta do primeiro grupo de frases, porém, de ambientes de chats das seguintes redes sociais: “Facebook”, “Twitter” e “Orkut”. Este conjunto considera-se como grupo “dois” (2).

Essa diferenciação entre esses grupos deu-se pelo objetivo de se tentar descobrir se existe alguma diferença na linguagem utilizada entre esses dois ambientes.

Para avaliar o sistema em números, utilizou-se de duas medidas conhecidas para avaliar sistemas de recuperação de informação (SRI): “Revocação (*Recall*)” e “Precisão”. [CARDOSO, 2002] explica que precisão é a fração dos documentos já examinados que são relevantes. Já a revocação é a fração dos documentos relevantes observados dentre os documentos examinados. A Figura 15 apresenta a ideia dessas duas medidas.

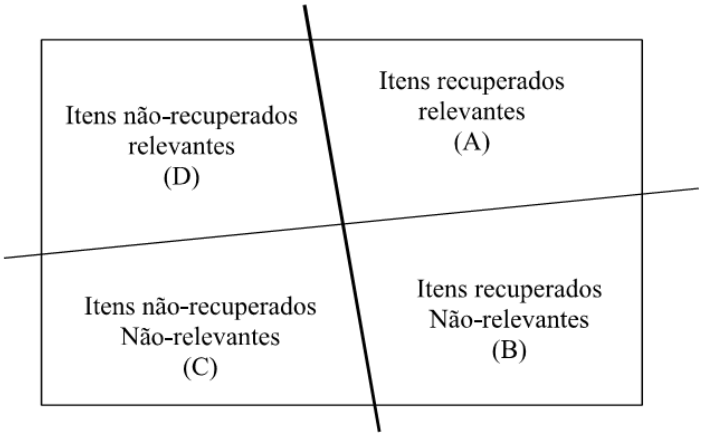

Figura 15: Figura utilizada como exemplo para demonstrar como se calcula os valores de precisão e revocação.

Para colocar valores nas duas medidas, apresentam-se as duas respectivas equações de acordo com os grupos apresentados na Figura 15.

$$Revocação = \frac{Grupo_A}{Grupo_A + Grupo_D} \qquad Precisão = \frac{Grupo_A}{Grupo_A + Grupo_B}$$

Outra medida que foi utilizada para avaliar o sistema foi o cálculo de F1 (média harmônica entre a precisão e a revocação). A métrica F1 faz uma ponderação da precisão e da revocação, podendo ser calculada pela fórmula abaixo:

$$F1 = \frac{2\ x\ Precisão\ x\ Revocação}{Precisão + Revocação}$$

É importante observar que as três métricas apresentadas são aplicadas por grupos de frases.

5.1. GRAU DE ACERTO DO SISTEMA

Abaixo são apresentados alguns dos testes do sistema que visam encontrar valores concretos para o grau de acerto do sistema. Os testes utilizam-se das medidas de Revocação, Precisão e F1.

5.1.1. EXPERIMENTO 1 – PRECISÃO, REVOCAÇÃO E F1 DOS MÓDULOS DE TRATAMENTO DE GÍRIAS E EXPRESSÕES

O primeiro experimento consiste em verificar qual o valor de revocação, precisão e F1 tanto do módulo de tratamento de gírias, bem como do módulo de

tratamento de expressões. Ao se calcular os valores de revocação procura-se saber quantas gírias e expressões foram coletadas pelos módulos de tratamento, de acordo com a quantidade de gírias e expressões que deveriam ser coletadas. Por exemplo: Se a base de dados contém apenas uma gíria (“vc”) e a frase possui duas gírias (“vc” e “tb”), o valor da revocação é de 0,5 ou 50%.

Já no caso dos valores de precisão identifica-se qual a porcentagem de gírias e expressões que foram corretamente tratadas de acordo com a quantidade total de gírias e expressões tratadas.

E por F1, interpreta-se uma média harmônica dos valores de precisão e revocação, onde quanto maior seu valor, mais congruente foi a atuação do sistema em cima dos grupos de frases.

Os resultados obtidos são mostrados na Tabela 1.

Módulo de Tratamento	Grupo de Frases	Revocação	Precisão	F1
Gírias	**1**	0,8973	0,9578	0,9266
Expressões	**1**	0,3378	0,8333	0,4808
Gírias	**2**	0,8732	0,9547	0,9122
Expressões	**2**	0,3421	1	0,5098

Tabela 1: Valores de Revocação, Precisão e F1 para os dois grupos de frases coletadas de acordo com os módulos de identificação e tratamento de gírias e expressões.

Analisando inicialmente os valores de revocação acima apresentados, é possível notar que, com grande disparidade, o sistema conseguiu captar uma quantidade muito maior de gírias do que expressões e regionalismos nas

mensagens dos usuários. Esses valores do módulo de tratamento de gírias, perto de 90%, podem ser atribuídos a uma base de dados bem preenchida em conjunto a uma boa captação das gírias.

Contudo, o módulo de tratamento de expressões apresentou resultados pouco esperados, com valores de 0,3378 para o grupo um de frases e 0,3421 para o grupo dois de frases. Entre os fatores que podem ter contribuído para esses valores, pode-se citar as pequenas divergências nos tempos verbais das expressões utilizadas. Por exemplo, a expressão: “ligado na parada” é diferente da expressão “se ligou na parada”. Isso indica que o sistema não esta apto a tratar casos assim.

Outro fator relevante seria o caso de que como o módulo de tratamento de gírias é executado primeiro que o módulo de tratamento de expressão pode-se ocasionar de que, uma palavra ou gíria dentro da expressão que seria verificada, ser tratada, implicando no não reconhecimento da expressão.

Verificando agora os valores de precisão, percebe-se que tanto o módulo de tratamento de expressão quanto o módulo de tratamento de gírias conseguiram corrigir da maneira correta mais que 80% em todos os casos.

O autor acredita que os valores poderiam ser maiores, se não fosse o fato de que usuários de chats estão em constante criação de novas gírias e expressões. E também ao fato de que uma mesma gíria pode apresentar vários significados, sendo isso, muito relativo de pessoa para pessoa.

Quanto aos valores de F1, apresentam-se congruentes aos valores de precisão e revocação, pois ao módulo de tratamento de gírias, teve-se uma grande quantidade de gírias encontradas e a grande parte delas, foi corretamente tratada. Já o módulo de tratamento de expressões, não conseguiu captar um bom número de expressões, implicando em um baixo valor de F1.

Considerando os valores encontrados no experimento, conclui-se que apesar de o sistema não estar captando bem expressões e regionalismos, ele apresentou resultados relativamente bons tanto ao tratamento de gírias quanto de expressões, apresentados pelo valor de precisão.

5.1.2. EXPERIMENTO 2 – REVOCAÇÃO DO MÓDULO DE TRADUÇÃO

Este segundo experimento tem por objetivo alcançar um valor de revocação para o módulo de tradução das mensagens do sistema. Este valor é, por si, um valor que representa a quantidade de palavras que o tradutor conseguiu traduzir, independentemente se o fez da maneira correta.

Torna-se importante ressaltar que avaliar tanto a precisão quanto a qualidade do módulo de tratamento é uma tarefa complicada, já que é totalmente relativo a um usuário dizer se uma tradução é boa ou não. Além disso, usuários podem divergir quanto a questão do que é realmente uma boa tradução.

Adicionando-se a este fator, tem-se o critério API de tradução. Como o sistema é dotado de uma API que realiza esse tratamento, torna-se independente do autor, e da implementação, a qualidade e as traduções das mensagens em si.

Assim sendo, a Tabela 2 apresenta os valores de revocação analisados nos dois grupos de frases para o módulo de tradução.

Grupo de Frases	Média da Revocação
1	0,8835
2	0,8471

Tabela 2: Valores de média dos valores de revocação analisados pelo módulo de tradução para o grupo dois de frases.

Os valores acima apresentam que o módulo de tradução, ou para ser mais exato, a API de tradução do *Bing Translator* foi capaz de capturar em ambos os casos quase 90% das palavras.

Este é um número significativo diante de frases que podem possuir algumas gírias e expressões (as não captadas pelos módulos anteriores). Apesar do valor alto, ressalta-se que este valor não indica quantas dessas palavras a API conseguiu traduzir corretamente, e sim, quantas ela identificou.

Acredita-se que a grande questão para a qual este valor esteja alto é devido a alta capacidade do sistema em geral, tratar gírias e expressões nos seus módulos anteriores de maneira correta. Ou seja, a precisão dos módulos de tratamento de gírias e expressões tem implicado significativamente neste valor.

5.2. USO DA BASE DE DADOS

5.2.1. EXPERIMENTO 3 – PORCENTAGEM DO USO DA BASE DE DADOS

Este experimento tem como objetivo averiguar qual a porcentagem de uso das bases de dados de gírias e a de expressões.

Os valores obtidos neste experimento em conjunto ao experimento um

fornecem uma resposta de quanto o sistema está sendo utilizado, e também se o sistema o faz da maneira correta.

As bases de dados de gírias e expressões contam com respectivamente, 570 gírias e 368 expressões cadastradas. Assim sendo, a Tabelas 3 apresenta a quantidade de gírias e expressões que foram encontradas para os dois grupos de frases.

Base de Dados	Grupo de Frases	Quantidades Encontrada	Porcentagem em relação ao total
Gírias	**1**	277	48,6%
Expressões	**1**	9	2,45%
Gírias	**2**	206	36,14%
Expressões	**2**	3	0,82%

Tabela 3: Valores de quantidade de gírias e expressões encontradas e o quanto esse valor significa no total de gírias e expressões na base de dados para os grupos de frases um e dois.

Na Tabela 3, a terceira coluna, significa qual a quantidade de diferentes gírias ou expressões encontradas, não a quantidade total existentes.

Através dos valores recém-apresentados, conclui-se que é comum encontrar uma maior variedade de gírias em ambientes de chats em relação à redes sociais.

Analisa-se também que o uso de gírias é muito mais comum que o uso de expressões e regionalismos. Apesar de seus valores não terem alcançado nem metade da quantidade de gírias existentes na base de dados, o experimento 1 mostrou que foi possível detectar a maioria do casos, constatando que a base de

dados de gírias está congruente com a necessidade dos usuários.

Já os baixos valores de expressões que foram encontradas podem indicar que os usuários não têm utilizado de tantas expressões nas suas mensagens, ou pode ter havido um não correto preenchimento da base de dados, ou mesmo problemas na captação das expressões, como foi comentado no Experimento 1.

5.3. DISCUSSÃO DOS RESULTADOS

Para a avaliação deste trabalho, realizou-se três (3) experimentos, analisando-os em dois cenários diferentes. Estes cenários são caracterizados pelo lugar onde foram coletadas as frases para realização dos experimentos.

Na secção 5.1, foram detalhados os resultados obtidos pelos módulos de tratamento de gírias, de expressões e de tradução. Observa-se através destes resultados que separadamente os módulos de tratamento de gírias e de tradução são capazes de identificar quase todas as gírias e palavras, respectivamente.

Apesar do módulo de tratamento de expressão não ter conseguido captar uma quantidade esperada de expressões, ele e o módulo de tratamento de gírias conseguiram tratar de maneira eficiente as gírias e expressões que foram encontradas.

Identifica-se um padrão de escrita mais erroneamente em ambientes de chats do que de redes sociais, apesar de ambos apresentarem quantidade relativamente grande de uso de gírias. Quanto ao uso de expressões, verificou-se que nem todas elas são utilizadas, contudo as que são, foram bem tratadas.

Por fim, analisa-se que o sistema em si tem cumprido com seu papel, apesar de ser notória a necessidade de melhora em sua base de dados para expressões.

6. CONCLUSÃO

Ao analisar a situação do problema proposto, constatou-se a real necessidade de criação de um sistema capaz de melhorar a comunicação entre as pessoas, de maneira a não afetar a coerência do texto, bem como a disposição das palavras.

Constatou-se ainda, que ambientes de chats apresentam formas de escrita rápidas e curtas, como [OTHERO, 2002] afirmou. Essa forma tem implicado no surgimento de expressões e vocábulos não oficiais na língua portuguesa.

Assim sendo, foi realizado com sucesso o objetivo de construir um sistema de chat *online* capaz de realizar a identificação, tratamento de gírias e expressões e, se necessária, a tradução das mensagens dos usuários para outro idioma.

Averiguou-se que este sistema poderá ser de grande valia quando aplicado a usuários com pouco conhecimento do idioma inglês e a usuários que utilizam de muitas gírias e expressões em suas comunicações, já que os experimentos mostraram que é possível recuperar e tratar esses casos.

Constatou-se ainda que o módulo de tratamento de gírias apresentou um bom índice de captura e tratamento das frases. Diferente deste, o módulo de tratamento de expressões apresentou um baixo índice de captura de expressões e regionalismos. Porém, esses resultados podem ser melhorados através da utilização de melhores métodos e técnicas de coleta e identificação, já que a forma escrita da expressão depende exclusivamente do usuário.

O módulo de tradução de idiomas, através de sua API, realizou sua função esperada, conseguindo traduzir a maior parte das palavras nas frases. O

autor acredita que um dos motivos que ajudou nesse índice foi o módulo de tratamento de gírias que conseguiu tratar a maior parte das gírias, evitando possíveis palavras desconhecidas para a API.

6.1. DIFICULDADES ENCONTRADAS

Um dos problemas encontrados na execução do trabalho ocorreu durante a coleta de gírias e expressões para a base de dados, já que isso depende exclusivamente dos usuários que utilizam esses sistemas. Muitas vezes, a maioria das frases encontradas apresentava as mesmas gírias e com os mesmos significados. Congruente a este problema, observou-se o mesmo na coleta de frases para análise dos módulos de tratamento e da base de dados.

Outro problema encontrado aconteceu no momento de conferir a correção dos dados que o sistema havia emitido, sendo necessário avaliar manualmente frase por frase captada, analisando e conferindo seus respectivos valores.

6.2. TRABALHOS FUTUROS

É possível o desenvolvimento ou integração de um analisador semântico ao sistema, para tratar casos que apresentam a razão de uma gíria com seu significado de um para mais de um. Nestas situações o analisador semântico poderia proporcionar a melhor expansão ou tratamento para a gíria ou expressão, devolvendo o valor correto, diferente do estado atual, que o sistema retorna todas as opções ao usuário.

É possível ainda implementar um módulo onde é possível ao usuário poder processar uma mensagem recebida de outro usuário, ou mesmo o sistema

sugerir a adição de gírias ou expressão, em palavras encontradas nas mensagens dos usuários, que sejam desconhecidas e que não estejam na base de dados.

Por fim, é possível ainda, a integração desse sistema para telefonia móvel, já que esse ambiente de conversação também apresenta um grande número de usuários que se comunicam com pessoas de outros países e culturas através de celulares e tablets.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ANACLETO, J. C.; CASELI, H. M.; FELS, S.; SUGIYAMA, B. A.. Using cultural knowledge to assist communication between people with different cultural background. SIGDOC '10 Proceedings of the 28th ACM International Conference on Design of Communication. New York, USA. 2010.

ARANHA, C. N.. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

ATALLAH, M. J.. *Algorithms and Theory of Computation Handbook. Appearing in the Dictionary of Computer Science, Engineering and Technology. CRC Press LLC.* 1999.

BECK, D. E.. Aprimorando o tratamento de Expressões Multipalavras em um tradutor automático baseado em regras. Trabalho de Conclusão de Curso. Bacharelado em Ciência da Computação, Universidade Federal de Rio Grande do Sul, Porto Alegre, RS, 2009.

BULEGON, H.; MORO, C. M. C.. Mineração de texto e o processamento de linguagem natural em sumários de alta hospitalar. *Text mining and natural language processing in discharge sumaries. J. Health Inform. Abril-Jun, 2(2):51-6.* 2010.

BUYYA, R.; SELVI, S. T.; CHU, X. *Object Oriented Programming with Java: Essentials and Applications*. Tata McGraw Hill Education Private Limited. June, 2009.

CARDOSO, O. N. P.. Recuperação de Informação. INFOCOMP. *Journal of Computer Science*. 2:33-38. 2000.

CARRILHO, J. . *Desenvolvimento de uma Metodologia para Mineração de Textos*. Dissertação de Mestrado, Departamento de Engenharia Elétrica, PUC-Rio. 2007.

CASELI, H. de M.; SUGIYAMA, B. A.. ANACLETO, J. C.. Using Common Sense to generate culturally contextualized Machine Translation, In *Proceedings of the NAACL HLT 2010 Young Investigations Worshop on Computational Approaches to Languages of the Americas,* Los Angeles, California, pp. 24-31. Junho, 2010.

DIETZEL, S.. Example-based Machine Translation. GRIN Verlag. Druck und Bindung: Books on Demand GmbH, Norderstedt Germany. Auflage, 2007.

FREITAG, R. M. K.; FONSECA E SILVA, M. Uma análise sociolinguística da língua utilizada na internet: implicações para o ensino da língua portuguesa. Revista Intercâmbio, volume XV. São Paulo: LAEL/PUC-SP, 2006.

GONÇALVES, T.; SILVA, C.; QUARESMA, P.; VIEIRA, R. *Analyzing Part-of-Speech for Portuguese Text Classification. Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*, (pp. 551-562). México City, México. 2006.

HAVASI, C.; ALONSO, R. S. J. B.. *ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. Proceedings of Recent Advances in Natural Language Processing.* 2007

Hutchins, W. J.; Somers, H. L.. *An introduction to machine translation*. *Academec Press, Harcourt Brace Jovanovich*. Publishers, London, PP. 259-278. 1992.

JUNG, C. F. Metodologia para pesquisa & desenvolvimento: aplicada a novas tecnologias, produtos e processos. Rio de Janeiro/RJ: Axcel Books do Brasil Editora, 2004.

Kit, C.; Wong, T. M.. *Comparative Evaluation of Online Machine Translation Systems with Legal Texts*. Law Library Journal, 100(2):299. 2008.

KONDACHY, M.. *Text Mining Application Programming*. Charles River Media. 2006.

LOPEZ, A.. *Statistical Machine Translation. University of Edinburgh.* ACM Computing Surveys, Vol. 40, No. 3, Article 8, August 2008.

MENEZES, A.; QUIRK, Chris.. *Microsoft Research Treelet Translation System*: *IWSLT Evaluation*. Proceedings of the International Workshop on Spoken Language Translation. October, 2005.

Minsky, M.. *The Society of Mind*. Simon and Schuster, New York. 1986.

NUNES, M. G. V.. A Construção de um Léxico da Língua Portuguesa do Brasil para suporte à Correção Automática de Textos. Relatórios Técnicos do ICMC-USP 42. LOE. ICMC. Setembro 1996.

OLIVEIRA, F. A. D. de.. Processamento de linguagem natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa. Revista de Ciência da Informação. Rio de Janeiro. n. 5. Maio 2002.

OTHERO, G. de Á.. *A língua portuguesa nas salas de b@te-p@po: uma visão lingüística de nosso idioma na era digital*. Novo Hamburgo: Editora do autor: 2002.

PAPIPENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.. *Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 311-318. July 2002.

SILVA, B. C. D.; MONTILA, G.; RINO, L. H. M.; SPECIA, L.; NUNES, M. das G. V.; OLIVEIRA, O. N.; MARTINS, R. T.; PARDO, T. A. S.. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. Núcleo Interinstitucional da Lingüística Computacional: Agosto de 2007.

SILVA, F. da. Análise Comparativa dos Resultados de Mecanismos de Tradução Automática Baseados em Regras e Estatísticas. Dissertação submetida à Universidade Federal de Santa Catarina para obtenção de grau de mestre em Estudos da Tradução. USFC, 2010.

SOARES, F. de A.. Mineração de textos na Coleta Inteligente de Dados na Web. Dissertação (Mestrado em Engenharia Elétrica). Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

SUGIYAMA, B. A.; ANACLETO, J. C.; CASELI, H. de M.. *Assisting users in a cross-cultural communication by providing culturally contextualized translations.* SIGDOC'11, Pisa, Italy. Proceedings of the 29th ACM international conference on Design of communication. Pg. 189-194. October 3–5, 2011

TRIPATHI, S.; SARKHEL, J. K.. *Approaches to Machine Translation. Annals of Library and Information Studies*. Vol. 57, pp. 388-393. , December 2010.

VAUQUOIS, B.. *A survey of formal grammars and algorithms for recognition and transformation in mechanical translation*. IFIP CONGRESS. p. 1114-1122. 1968.

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

YUJIAN, L.; LIU, B.. *A Normalized Levenshtein Distance Metric.* IEEE Transactions Pattern Analysis and Machine Intelligence. Vol. 29, p. 1091 – 1095. June, 2007.