UNIVERSIDADE DO SUL DE SANTA CATARINA

JOÃO PEDRO PONTES MARTINS

PROPOSTA DE IMPLEMENTAÇÃO DE UM CHATTERBOT COM ANÁLISE DO HISTÓRICO DA CONVERSA PARA REALIZAR A DESAMBIGUAÇÃO LÉXICA DE SENTIDO

Palhoça

2013

JOÃO PEDRO PONTES MARTINS

PROPOSTA DE IMPLEMENTAÇÃO DE UM CHATTERBOT COM ANÁLISE DO HISTÓRICO DA CONVERSA PARA REALIZAR A DESAMBIGUAÇÃO LÉXICA DE SENTIDO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade do Sul de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Aran Bey Tcholakian Morales, Dr.

Palhoça

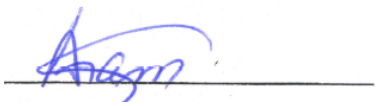
2013

JOÃO PEDRO PONTES MARTINS

PROPOSTA DE IMPLEMENTAÇÃO DE UM CHATTERBOT COM ANÁLISE DO HISTÓRICO DA CONVERSA PARA REALIZAR A DESAMBIGUAÇÃO LÉXICA DE SENTIDO

Este Trabalho de Conclusão de Curso foi julgado adequado à obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo Curso de Graduação em Ciência da Computação da Universidade do Sul de Santa Catarina.

Palhoça, 18 de junho de 2013.



Professor e orientador Aran Bey Tcholakian Morales, Dr.
Universidade do Sul de Santa Catarina

Prof. Maria Ines Castiñeira, Dr.
Universidade do Sul de Santa Catarina

Prof. Andréa Bordin, Msc.
Universidade do Sul de Santa Catarina

RESUMO

Este trabalho trata da problemática da ambiguidade léxica de sentido em sistemas chatterbots. Decidiu-se como objetivo geral a construção de um chatterbot que realize o tratamento de desambiguação léxica de sentido (DLS) numa frase digitada pelo usuário. Para isso argumentou-se que com o crescimento da internet e da utilização dos computadores faz-se interessante uma interação homem-máquina mais natural, optando assim pela conversa em linguagem natural. A revisão bibliográfica apresenta um histórico dos chatterbots e suas técnicas de implementação, e na área de Processamento de Linguagem Natural (PLN) os métodos e fontes de conhecimento utilizados para realizar a DLS. Este trabalho se enquadra em pesquisa aplicada, exploratória e bibliográfica. No desenvolvimento do chatterbot aplicou-se técnicas da DLS, utilizando como fontes de conhecimento: *bag-of-words*, collocations, associação semântica de palavras e *stoplist* para atingir os objetivos levantados, já a interação direta entre o sistema e o usuário foi realizada através da técnica de casamento de padrão simples. Os resultados obtidos através da validação com os usuários foram altamente satisfatórios, o sistema chatterbot conseguiu realizar a desambiguação em grande parte das perguntas, se mostrando eficaz para solucionar a problemática levantada.

Palavras-Chave: Chatterbot. Inteligência Artificial. Processamento de Linguagem Natural. Desambiguação Léxica de Sentido.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Alan Turing, em 1950, com sua publicação "*Computing Machinery and Intelligence*", levantou o seguinte questionamento: "As máquinas podem pensar?", e através do famoso "Teste de Turing", apresentado sob o título "*The Imitation Game*", em seu artigo, propôs uma avaliação para identificar se um computador poderia ser chamado de inteligente (TURING, 1950). Esse teste, segundo Coppin (2012) baseia-se na ideia de que se uma pessoa interrogasse um computador e não pudesse dizer se este era mesmo um computador ou um ser humano, então, para todos os efeitos, o computador poderia ser chamado de inteligente.

A partir da publicação do artigo de Turing, muito foi estudado na área de processamento de linguagem natural e as primeiras tentativas de construir um computador que compreendesse a linguagem humana foram sendo desenvolvidas. (COPPIN, 2012).

Durante o passar dos anos, muitos sistemas foram construídos tendo como propósito manter uma conversa com um ser humano, como é o caso do sistema ELIZA. “ELIZA, o primeiro sistema que propunha conversar com usuários em linguagem natural, foi projetado por Weizenbaum (1966) seguindo as ideias de Turing (1950)” (NEVES; BARROS, 2005, p.1032). Os mesmos autores afirmam que o ELIZA foi o primeiro e o que abriu as portas para a possibilidade real do desenvolvimento de “máquinas de conversar”, hoje conhecidas como chatterbots (robôs de conversação).

Ao falar do chatterbot ELIZA, Primo e Coelho (2002, p. 3) afirmam que “O objetivo desse pequeno programa, de apenas 204 linhas de código, é simular uma conversação entre uma psicóloga de estilo rogeriano e seu paciente”.

Para Laven (2012), um chatterbot é um programa de computador que procura simular uma conversação com o usuário, com o objetivo de fazê-lo pensar temporariamente que está conversando com outro ser humano.

Hoje existe uma gama de chatterbots muito interessantes, cada um com suas peculiaridades e características, sendo utilizados nas mais diversas áreas, e executando as mais diversas funções. Dentre os chatterbots mais famosos, além do ELIZA, pode-se citar o Profª Elektra, desenvolvido na UFRGS com o intuito de responder perguntas sobre Física para alunos do ensino médio em preparação para o vestibular, e que posteriormente foi alterado para cursos a distância. (LEONHARDT et al. 2003); o Cybelle que é o primeiro chatterbot na web a falar português, com uma estrutura semelhante a do ELIZA, e que em seu site apresenta uma interface gráfica feminina, e um arquivo de áudio com a “voz” do robô. (PRIMO;

COELHO, 2002); e o ALICE que ganhou o Loebner Prize[1] em 2000, 2001 e 2004, e foi considerado o "computador mais humano" que já concorreu na premiação (WALLACE, 2012).

Posteriormente, neste trabalho, será mostrado um histórico dos chatterbots com a sua evolução.

1.1 PROBLEMÁTICA

Com décadas de estudos e pesquisas nas áreas de inteligência artificial e processamento de linguagem natural, os chatterbots evoluíram muito desde o artigo publicado por Alan Turing em 1950, porém ainda existem muitas dificuldades e obstáculos em seus projetos e implementações.

Allen (1995, apud NEVES; BARROS, 2005 p. 1032) afirma "Chatterbots podem ser vistos como aplicações de Processamento de Linguagem Natural (PLN), sofrendo, portanto, de problemas comuns a esses sistemas, tais como ambiguidade léxica e semântica".

Segundo Specia e Nunes (2004) a ambiguidade léxica ocorre da necessidade de escolha de um dos sentidos de uma palavra numa frase, sendo causada principalmente pela polissemia e homonímia.

Nesse trabalho será focado na ambiguidade léxica, que segundo Coppin (2012), acontece quando para uma mesma palavra existe mais de um significado.

Apesar de ser visto como uma aplicação de PLN, ele engloba também outros fatores e requisitos, que fazem com que existam dificuldades específicas deles, como afirma Leonhard (2003, p.2) sobre o ELIZA " [...] Não há uma memória no robô, ou seja, ela não consegue lembrar o que foi falado anteriormente."

Correa (2011, p. 52) afirma que "Um dos problemas que pode ocorrer em sistemas de conversação, como chatterbot, é a limitação deste, de manter o contexto da conversação, sem desviar do assunto ou tornar a repetir certas frases".

Com esses argumentos, vê-se que ainda há muito a se estudar sobre os chatterbots, suas teorias e técnicas para implementação. Dois assuntos muito abordados na bibliografia

[1] Premio Loebner: Concurso que premia o computador em que suas respostas sejam indistinguíveis das de uma pessoa.

consultada são a questão de manter uma memória para o robô, para haver um histórico da conversa, e o tratamento de ambiguidade léxica e semântica, que se aplica ao processamento de linguagem natural.

Após essa explanação, questiona-se a possibilidade de um chatterbot guardar o histórico de conversação e, através dele, fazer a desambiguação de frases.

1.2 OBJETIVOS

A seguir, serão apresentados os objetivos deste trabalho.

1.2.1 Objetivo Geral

Este trabalho tem por objetivo geral propor a construção de um chatterbot com desambiguação léxica de sentido, utilizando como contexto para análise da palavra o histórico da conversa corrente.

1.2.2 Objetivos Específicos

Entre os objetivos específicos pode-se citar:

- analisar a anatomia de um chatterbot;
- pesquisar o histórico dos chatterbots;
- analisar as técnicas de implementação utilizadas nos chatterbots;
- estudar os métodos de desambiguação léxica de sentido.
- construir um protótipo de um chatterbot com desambiguação léxica de sentido.

1.3 JUSTIFICATIVA

As pesquisas na área dos chatterbots foram um pouco esquecidas devido ao fato de, na época de seu auge, não haver um ambiente propício para o seu desenvolvimento, contudo a popularização e o crescimento da internet acabaram reavivando seus trabalhos, e seu desenvolvimento acabou ganhando mais espaço, com isso os chatterbots ganharam mais reconhecimento e hoje são utilizados por milhares de pessoas (NEVES; BARROS, 2005).

Jacob Junior (2008) afirma que chatterbots podem ser utilizados em aplicações que façam suporte ao consumidor, aplicações de ensino a distância, marketing, faq e etc.

Moura (2008, p.15) também demonstra que, "apesar dos chatterbots existirem a décadas, o crescente aumento na utilização dessa tecnologia ocorreu a partir da popularização da Internet. A Internet fez os chatterbots acessíveis ao público em geral".

Cada vez mais está se tornando essencial que os computadores entendam linguagens naturais, tanto para análise de dados textuais não estruturados espalhados pela internet, quanto para interação com o ser humano. A ideia de trazer para as pessoas a possibilidade de realizar pesquisas na internet, utilizando sua própria linguagem, recuperar informação de texto e fazer tradução automática tem se tornado muito popular (COPPIN, 2012).

Os chatterbots têm se mostrado uma excelente ferramenta para interação do computador com o ser humano, incentivando as pesquisas na área de processamento de linguagem natural. Para que essa interação ocorra da melhor forma possível, as respostas obtidas dos chatterbots precisam ser coerentes e relevantes para o que o usuário deseja, perante isso se observa que a utilização de um modelo de representação de conhecimento, como uma ontologia, enquadra-se perfeitamente no requisito de melhorar o mapeamento das respostas (CAFÉ; COMARELLA, 2008).

Um dos fatores mais importantes na comunicação é o que diz respeito ao significado de uma frase que pode expressar um conhecimento de mundo ou uma intenção do falante para o ouvinte. Para desenvolver um sistema que faça essa distinção na frase, que consiga fazer uma desambiguação de sentido, é necessário recorrer a técnicas de representação de conhecimento, incluindo algoritmos que façam a relação entre os componentes de um texto para realizar a desambiguação (VIEIRA; LIMA, 2001).

Devido a essa necessidade e a ausência de material bibliográfico sobre chatterbots tratando palavras ambíguas, esse trabalho propõe a implementação de um chatterbot com desambiguação léxica de sentido.

Dessa forma, vê-se a importância de unir os estudos na área da linguística com os métodos computacionais disponíveis para resolver essas questões que ainda deixam brechas nos sistemas chatterbots, tornando-os mais frágeis em sua conversação, e, através disso, evoluindo nos trabalhos referentes a essa área, que podem ser utilizados nas mais diversas aplicações de chatterbots e ao processamento de linguagem natural, como na educação e no entretenimento.

Perante isso vê-se a necessidade de uma melhor comunicação, utilizando linguagem natural entre homem e máquina, visando a construir não apenas sistemas capazes de manter uma conversa com melhor fluência e coerência, mas fazendo dessa maneira a aproximação do computador com o usuário, tornando a interação com a máquina uma atividade mais humana, confortável e agradável.

1.4 ESTRUTURA DA MONOGRAFIA

No primeiro capítulo, consta a introdução do trabalho com uma apresentação do tema, os objetivos gerais e específicos, a problemática e a justificativa.

No segundo capítulo, encontra-se a revisão bibliográfica, contendo todo o embasamento teórico necessário para o tema abordado, sendo estudado o histórico de chatterbots, suas técnicas de implementação, a problemática envolvida em sua construção, a questão da ambiguidade de uma palavra e de como resolvê-la e os estudos da teoria linguística e computacional necessários para alcançar os objetivos do trabalho.

No terceiro capítulo, é apresentado o método, a caracterização do tipo de pesquisa, as etapas metodológicas e as delimitações para este projeto.

No quarto capítulo, é proposta a solução para um chatterbot com desambiguação léxica de sentido, apresentando as etapas a se percorrer para atingir os objetivos.

No quinto capítulo, a conclusão fecha o trabalho apresentando os obstáculos encontrados durante o desenvolvimento do chatterbot e trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo visa a dar todo o embasamento teórico necessário sobre os chatterbots, seu histórico, suas técnicas de implementação, as dificuldades e os obstáculos enfrentados no seu desenvolvimento.

2.1 HISTÓRICO DOS CHATTERBOTS

A ideia de chatterbot não é nova, muito já foi produzido nessa área, e cada um dos programas desenvolvidos tiveram sua importância e incentivo para a formulação do próximo.

Rothemel (2007) identifica três gerações de chatterbots: A primeira utiliza-se de casamento de padrões e regras gramaticais, e tem, como exemplo, o Eliza. Essa geração não tem memória, pois não mantém um histórico das conversas anteriores. A segunda geração pode ser representada pelo chatterbot JULIA, que baseia-se em técnicas de inteligência artificial como as regras de produção e redes neurais, e a terceira geração, que é a mais recente, consiste em utilizar linguagens de marcação como AIML para sua base de conhecimento e tem como principal representante o robô ALICE.

2.1.1 ELIZA

ELIZA, considerado um dos chatterbots mais antigos, foi desenvolvido no *Massachussets Institute of Technology* (MIT), em 1966, pelo professor Joseph Weizenbaum, e tinha como função exercer o papel de um psicanalista conversando com seu paciente. Eliza seguia o estilo da terapia rogeriana, interagindo com o paciente em forma de perguntas para estimulá-lo a refletir sobre suas próprias emoções (LEONHARDT, 2005).

Esse chatterbot foi muito além de apenas um programa de computador, segundo Primo e Coelho (2002, p.3) “Eliza é um dos programas de Inteligência Artificial mais antigo e mais conhecido no mundo. Pode-se, também, dizer que é um dos programas mais estudados na história da informática”.

O chatterbot Eliza utiliza-se de palavras chaves que se baseiam em regras de decomposição, que fazem a separação da frase digitada pelo usuário e, através de palavras da própria pergunta, é montada a resposta e, além disso, leva-se em consideração também o contexto da palavra. Com isso conclui-se que o Eliza é baseado em palavras-chave e em reestruturação da pergunta do usuário (LEONARDT, 2005).

A importância desse programa vai além das técnicas e de utilizações de algoritmos, como relata Canuto (2005) “Esse sistema despertou atenção na comunidade científica, pois sistemas como esses, relativamente simples de serem implementados, são capazes de influenciar o comportamento dos usuários”. Ao ser lançado, Eliza surpreendeu a todos, psiquiatras acreditavam que poderia ser construído um robô de conversação totalmente automatizado para a psiquiatria e os usuários acabaram se envolvendo muito rapidamente pelo Eliza, a própria secretária de Weizenbaum pediu para ser deixada a sós por um momento com o chatterbot (PRIMO; COELHO, 2002 p. 3).

Além de surpreender o público de usuários, os psiquiatras e muitos cientistas da computação, o próprio Weizenbaum ficou espantado com a reação positiva do público. O motivo de todo esse sucesso do Eliza se deve ao fato do estilo de respostas, que são com sentenças interrogativas, isso faz com que o usuário enxergue significados nas curtas palavras do chatterbot, que foram digitadas pelo próprio usuário e que tolere respostas evasivas. (HUTCHENS; 1999, apud PRIMO; COELHO, 2002 p. 3).

Eliza, porém, têm algumas limitações. Primo e Coelho (2002) afirmam que, apesar de ter uma das personalidades mais bem definidas entre os robôs de conversação, de ser uma aplicação pioneira, e de ser bastante compreensiva com seus pacientes, Eliza não se lembra de nada do que foi dito na interação com o usuário. Leornhardt (2005), também, fala das limitações de Eliza, quando comenta que, além da ausência de memória, a formação de algumas frases através de sua técnica de modificação da sentença gera, algumas vezes, diálogos muito confusos e observa que o robô produz frases sem sentido se a entrada do usuário apresenta complementos em sua estrutura normal.

Abaixo segue uma imagem de uma conversa do Eliza com um usuário.

Figura 1 – Exemplo de conversa com o chatterbot Eliza.

```
User: Men are all alike.
Eliza: IN WHAT WAY?
User: They're always bugging us about something or other.
Eliza: CAN YOU THINK OF A SPECIFIC EXAMPLE?
User: Well, my boyfriend made me come here.
Eliza: YOUR BOYFRIEND MADE YOU COME HERE
User: He says I'm depressed much of the time.
Eliza: I AM SORRY TO HEAR YOU ARE DEPRESSED
User: It's true. I am unhappy
Eliza: DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE  UNHAPPY
User: I need some help, that much seems certain.
Eliza: WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
User: Perhaps I could learn to get along with my mother.
Eliza: TELL ME MORE ABOUT YOUR FAMILY
User: My mother takes care of me.
Eliza: WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
User: My father.
Eliza: YOUR FATHER
User: You are like my father in some ways.
```

Fonte: Weizenbaum (1966, p. 37).

2.1.2 PARRY

Continuando na linha do tempo dos chatterbots, temos o paranóico Parry, criado por Kenneth Colby, em 1971. Esse robô de conversação, através de um rastreamento de seu estado emocional interno, tem como principal objetivo simular um comportamento paranoico (LEONHARDT, 2005). Parry pode se enquadrar juntamente com o Eliza dentro da categoria dos chatterbots clássicos, que foram os primeiros a serem desenvolvidos com o objetivo inicial de estudar a complexidade na comunicação em linguagem natural entre homens e máquinas (ROTHERMEL; DOMINGUES, 2007).

Um dos pontos mais importantes nesse chatterbot, pelo fato de seu objetivo ser simular uma atitude paranoica, é seu aspecto emocional. Seu modelo foi projetado para simular comportamentos como medo, raiva e desconfiança e, com isso, classificar o usuário em amigável, hostil ou dominante. Contudo não obteve o mesmo sucesso que o Eliza, porém, devido a todos estes aspectos relatados, sua implementação estende os algoritmos utilizados no Eliza (WILKS; CATIZONE, 2000 apud. LEONHARDT, 2005).

O algoritmo para fazer esse tratamento de emoção funciona da seguinte forma: Parry têm variáveis como ‘raiva’ e ‘medo’ que são incrementadas ou decrementadas de acordo com o texto digitado pelo usuário, e as respostas são selecionadas de acordo com as alternativas disponíveis para a entrada do usuário (WILKS; CATIZONE; TURUNEN, 2006, p. 5).

Segue, abaixo, uma conversa do chatterbot Parry.

Figura 2 – Exemplo de conversa com o chatterbot Parry

```
User: Good afternoon. My name is Doctor Jones.
Parry: Pleased to meet you.
User: How are you ?
Parry: all right I guess.
User: Could you tell me your name ?
Parry: Pat Smith.
User: How old are you ?
Parry: 28.
User: I thought you might have some problems.
Parry: I've been upset lately.
User: By what ?
Parry: People get on my nerves sometimes.
User: Tell me more.
Parry: Do you know anything about bookies ?
User: No. What about bookies ?
Parry: A bookie didn't pay me off once.
User: When did this happen ?
Parry: A couple of months ago.
```

Fonte: Colby (1981, apud Hutchens, 1997).

2.1.3 JULIA

Em 1990, com os avanços nas áreas de inteligência artificial, processamento de linguagem natural e engenharia de software, os cientistas começaram a desenvolver sistemas, utilizando as novas técnicas disponíveis, como as redes neurais, dando início a segunda geração de chatterbots, como principal representante se pode citar o JULIA (CANUTO, 2005).

JULIA foi desenvolvido por Michael Mauldin na *Carnegie Mellon University*, em 1994, para habitar no jogo *TinyMUD*, um jogo em rede no qual vários usuários interagem via

teclado (MAULDIN, 1994), e sua função era auxiliar e fornecer dicas para os outros personagens (JACOB JUNIOR, 2008).

Pelo fato do TinyMUD propiciar um mundo com diversas pessoas se comunicando via teclado, foi o momento perfeito para por em prática o teste de Turing, pois todos os jogadores acabariam por serem juízes e entrevistadores em potencial (MAULDIN, 1994). Este robô de conversação tem uma certa vantagem em relação aos seus antecessores, como relata Jacob Junior (2008, p. 14): "Diferente dos outros chatterbots, JULIA possui bastante conhecimento sobre o mundo onde atua, uma vez que este tem um escopo reduzido, fornecendo um nível de conversação aceitável sobre o assunto".

Em sua primeira versão, Julia apresentava um algoritmo simples, do tipo "if-then-else", contudo, com o decorrer do tempo, suas habilidades na conversação foram alteradas para um modelo mais robusto, baseado em redes neurais (LEONHARDT, 2005).

As técnicas utilizadas no JULIA se utilizam basicamente de redes neurais para criar seu modelo de conversação (CANUTO, 2005), regras de casamento de padrões denominadas pelo próprio Mauldin de Redes Conversacionais (JACOB JÚNIOR, 2008).

Para explicar como essas redes funcionam, Leonhardt (2005, p. 27) relata:

> Dentro da rede, cada nó consiste em um conjunto de padrões, uma resposta simples, uma lista de nós estimulados e outra de nós inibidos. Quando a entrada do usuário aciona algum padrão, os nós que contêm o modelo têm sua ativação estimulada e o de maior nível é selecionado, enviando sua resposta ao usuário.

Julia, semelhante ao Parry, também agrega uma personalidade, porém essa é uma personalidade mais feminina, num determinado período do mês, Julia afirma estar com tensão pré-menstrual e, inclusive, seu humor pode se alterar (LEONHARDT, 2005).

Abaixo, segue um trecho da conversa do Julia no prêmio Loebner de 1993 (CHEONG, 1996 apud JACOB JUNIOR, 2008).

Figura 3 – Exemplo de conversa do chatterbot Julia com um juiz do prêmio Loebner.

Judge: Do you have a pet at home ?
Julia: *I'd rather have a real pet, like a cat. Cats are my*
favourite, because you can pet cats and they're easy to keep.
Judge: Do you know anything about miniatures schnauzers ?
Julia: *I don't know anything about miniature schnauzers, Judge.*
Judge: Well, I have one. We call him Pucci. He is
real a crazy cat - have you ver had a cat ?
Julia: *What a coincidence! I have two cats.*
Judge: I prefer dogs myself. Cats are too independent and
they always jump up on. By the way, are you a hep cat ?
Julia: *People don't own cats... it's the other way around*

Fonte: Cheong (1996, apud Junior, 2008).

2.1.4 Barry e Alfred

Da segunda geração, dentre outros chatterbots a serem comentados, encontra-se o Barry e o Alfred, que utilizam da tecnologia FRED criada por Robby Garner na Robitron Software Research (LAVEN, 2012), que possibilitou a construção de chatterbots para a web com a linguagem Java, o JFRED. Segundo Jacob Júnior (2008), a base de conhecimento dos sistemas criados com essa tecnologia se utiliza da linguagem JRL (*Java Ruleset Language*), que é baseada em lógica difusa para a construção de regras para o mapeamento das palavras-chaves e expressões regulares a serem utilizadas no casamento de padrões das frases digitadas pelo usuário.

Laven (2012) afirma que FRED é um experimento para explorar a utilização de linguagens naturais na interação entre humanos e máquinas, em particular, o estudo de conversação entre pessoas e maneiras de como os programas de computador podem aprender com as conversas dos humanos para obterem sua própria conversação.

FRED, apesar de pertencer à segunda geração de chatterbots diante de suas técnicas de inteligência artificial, como a lógica difusa, ele está relacionado junto com Eliza e Julia aos chatterbots clássicos, pois fazem parte dos robôs de conversação cujo objetivo é estudar a comunicação em linguagem natural entre homem e máquina. (ROTHERMEL, 2007, p. 3).

2.1.5 ALICE

Na terceira geração, tem-se os chatterbots desenvolvidos, utilizando linguagem de marcação AIML e, como principal representante desta geração, o ALICE (*Artificial Linguistic Internet Computer Entity*) criado por Richard S. Wallace em 1995 na *Lehigh University* (MOURA, 2008). A linguagem AIML permite as pessoas adicionarem conhecimento em chatterbots baseados na tecnologia do ALICE, essa linguagem foi desenvolvida durante os anos de 1995 e 2000 e foi a base do primeiro Alicebot (WALLACE, 2012).

O ALICE é um dos chatterbots mais populares na atualidade e seu sucesso se deve a sua inovação: a forma de como foi apresentado, como demonstra Leonhardt (2005, p. 27) "[...] além de muita documentação, apresenta uma saudação sonora ao visitante. [...] tem um grande poder de comunicação, além de uma interface gráfica que estimula o diálogo".

ALICE ganhou o Prêmio Loebner em 2000 e 2001, sendo que, no ano de 2001, obteve uma nota maior que a nota de um humano, fato inédito no concurso (LEITÃO, 2004).

2.1.6 Cybelle

No Brasil também há chatterbots conceituados, como o Cybelle, o primeiro chatterbot na web a falar português simulando um diálogo com o internauta com uma estrutura semelhante a do Eliza (PRIMO; COELHO, 2002).

Esse chatterbot consiste em duas partes, o mecanismo e o conhecimento, a relação entre elas é dada via lógica estímulo-resposta. A entrada do usuário é analisada, buscando por estímulos previstos ou suas combinações e essa análise obedece a critérios de relevância, associando respostas específicas sobre temas específicos, dessa maneira, assuntos como "futebol" e "esporte" exigem respostas diferenciadas, e mesmo ela podendo se dirigir ao usuário pelo nome e usar trechos das perguntas em suas respostas, ela não tem memória, ou seja, ela não "lembra" qual foi a pergunta anterior e não tem autonomia para montar sua própria resposta ao internauta (PRIMO; COELHO, 2002).

Quanto à variação de respostas para uma mesma entrada do usuário, Moura (2008, p. 35) afirma “Muitas são as situações em que para um mesmo estímulo podem ter mais de uma resposta prevista, possibilitando a variação entre essas alternativas, evitando assim a repetição”.

Cybelle, assim como os chatterbots Parry e Julia também, tem um pouco de personalidade, primeiramente pelo diferencial de ter uma ilustração feminina no site e um arquivo de áudio que cumprimenta o usuário dizendo “Oi! Meu nome é Cybelle”, que seria a “voz” do robô e, além disso, suas respostas demonstram uma certa “depressão” por “saber” de sua condição como robô (PRIMO; COELHO, 2002).

2.1.7 Robô ED

Desenvolvido especificamente para a Petrobrás, o robô ED é capaz de conversar com os usuários, como um atendente real, e falar sobre o uso racional de energia, derivados do petróleo, meio ambiente, gás natural, dicas de economia, qualidade do ar, biocombustíveis, programas educacionais e fontes alternativas de energia (MOURA, 2008).

2.1.8 Elektra

Criado na UFRGS e disponibilizado em 2002, o Elektra teve como objetivo inicial responder perguntas sobre física para alunos se preparando para o vestibular, contudo, posteriormente, em 2003 foi expandida sua base de conhecimento para agregar conceitos sobre redes de computadores para ser utilizado no curso de Especialização a Distância em Informática na Educação, na disciplina de Internet para Educadores (LEONHARDT, 2003).

2.2 TÉCNICAS DE IMPLEMENTAÇÃO

Quando o usuário utiliza o computador para efetuar alguma tarefa, o computador age como ferramenta e como co-participante da comunicação, pois o usuário entra com uma informação em um determinado formato entendível pelo computador e o computador, após processar essa informação, retorna a uma resposta entendível para o usuário (WAZLAWICK; CASTANHO, 2002).

Geralmente, essa informação entrada pelo usuário é feita através de cliques em botões ou comandos, porém essa não é a forma mais intuitiva e natural para se interagir.

Provavelmente, a forma mais natural para o ser humano se comunicar com outro ser humano seja através de expressões faciais, gestos, escritas, desenhos e outras modalidades de representação de informação. Perante isso, vê-se que uma forma de tornar a comunicação entre homem-máquina algo mais natural seria simulando diálogos via escrita, como os que o ser humano tem no dia-a-dia (BERSEN; 1998, apud WAZLAWICK; CASTANHO, 2002).

Para fazer a construção desses sistemas de conversação, Suereth (1997, apud WAZLAWICK; CASTANHO, 2002) afirma que existem dois tipos de interface que permitem a comunicação em linguagem natural entre homem e máquina: Os Processadores de Conversação e os Processadores de Linguagem Natural. O primeiro tipo é mais simples, seu conhecimento é diferente daquele que o ser humano necessita para exercer suas funções diárias e sua meta é simular uma conversação eficiente com o usuário, já o segundo necessita de estruturas para separar a informação em tipos organizados e, para isso, utiliza-se de analisadores léxicos, sintáticos, semânticos, de discursos e pragmática, incluindo mecanismos de máquina de inferência. Seu objetivo é “entender” a entrada do usuário e gerar novo conhecimento.

Este trabalho se enquadra no primeiro grupo, o de Processadores de Conversação, e para isso serão demonstrados alguns algoritmos e técnicas para fazer o tratamento desde a entrada digitada pelo usuário, o processamento executado pela máquina e a resposta advinda desse processo.

2.2.1 Casamento de Padrão

Segundo Wazlawick e Castanho (2002) e Leonhardt (2005), o casamento de padrões é uma técnica que consiste em um casamento entre um conjunto de palavras-chave e um grupo de respostas relacionadas àquelas chaves, respeitando a ordem das chaves.

Essa técnica é utilizada nos chatterbots de duas maneiras: Através do casamento de padrão simples, e através de uma linguagem própria de marcação, a AIML (JACOB JUNIOR, 2008).

2.2.1.1 Casamento de Padrão Simples

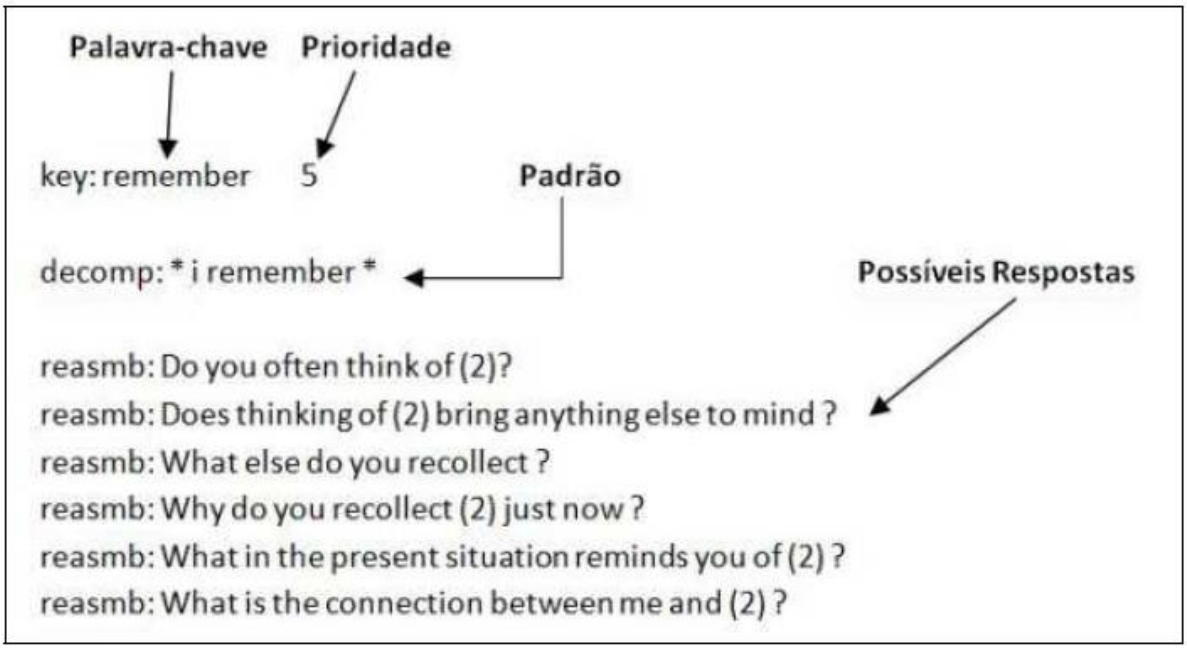
Esta técnica é utilizada nos chatterbots da primeira geração, como o Eliza e o Parry, aqui será demonstrado como funciona o algoritmo do Eliza:

O sistema identifica as palavras mais importantes digitadas pelo usuário e aplica uma regra de modificação na sentença, classificando a importância da palavra de acordo com um mini-contexto, como o sujeito, verbo, etc. (JACOB JUNIOR, 2008).

A Figura 1 demonstra um exemplo em que o usuário inseriu como entrada para o chatterbot a sentença "*I remember Dave*". Fazendo a análise deste processo, tem-se:

a) palavra-chave: *remember*
b) prioridade desta palavra-chave: 5
c) padrão Estabelecido: * *i remember* *
d) possíveis respostas para o padrão de entrada do usuário: "*Do you often think of (2) ?*", "*Does thinking of (2) bring anything else to mind ?*"

Figura 4 – Exemplo de técnica de processamento do Eliza



Fonte: Jacob Junior (2008, p. 9).

As respostas podem conter palavras digitadas pelo usuário para preencher os espaços representados na figura por (2). O (2) seria substituído pela palavra após "*remember*" (JACOB JUNIOR, 2008).

No exemplo citado anteriormente, em que, caso o usuário digitasse a sentença: "*I remember Dave*", a resposta do chatterbot para o padrão * *i remember* * poderia ser: "*Do you often think of Dave ?*", "*What in the present situation reminds you of Dave ?*" ou todas as respostas que estão presentes na Figura 1.

Em alguns chatterbots que implementam essa técnica de casamento de padrões, como o Parry, caso ele não possua conhecimento sobre o assunto citado, ou seja, não foram encontradas palavras-chave para seu assunto, é retornando ao usuário uma resposta sobre um tema específico de sua base de conhecimento (JACOB JUNIOR, 2008).

A utilização de técnicas com palavras-chave para a implementação de um robô de conversação é muito eficiente, têm-se utilizado o mesmo princípio dessa tecnologia tanto no Eliza através de algoritmos mais simples, quanto no ALICE, que utiliza de palavras-chave através de uma linguagem de marcação AIML (JACOB JUNIOR, 2008).

Juntamente com a ideia de palavras-chave e casamento de padrões, pode-se citar a utilização de **formas canônicas**. "A forma canônica é a diversidade de representação de uma pergunta, frase ou sentença. Entradas de perguntas distintas devem ser relacionadas a uma mesma representação de significado" (DELUCCA, 2002, apud OLIVEIRA et al, 2009).

Para melhor compreensão da forma canônica, pode-se citar, como exemplo, as seguintes sentenças: “A capa deste livro é azul?”, “Este livro tem a capa azul?”, “Azul é a capa deste livro?”.

Todas essas três perguntas dizem respeito se a capa do livro é azul, mas cada uma é escrita de uma maneira.

Segundo Delucca, 2002 (apud OLIVEIRA et al. 2009) é importante aplicar um processo para as formas canônicas para descobrir o assunto tratado e, através dessa descoberta, apresentar a resposta que faça mais sentido independentemente de como foi feita a pergunta. Para aplicar esse processo, pode-se utilizar de palavras-chaves para chegar ao assunto em questão, apresentando uma resposta coerente ao questionamento.

Dessa forma, conclui-se que, para responder as três questões, propostas anteriormente neste trabalho, e chegar ao assunto do qual elas tratam, pode-se utilizar de palavras-chave, e essas poderiam ser “capa”, “livro” e “azul”. Qualquer pergunta que o usuário digitasse como entrada, envolvendo essas três palavras estaria ligada ao mesmo assunto, retornando a mesma ideia de resposta.

Além disso, Eliza utiliza de saídas para não responder diretamente o questionamento do usuário, levando-o praticamente a conversar sozinho (LAUREANO, 1999).

2.2.1.2 Casamento de Padrão com Linguagem de Marcação (AIML)

A AIML é baseada em módulos que são chamados de categorias, cada categoria possui um padrão de entrada que será comparado à sentença submetida pelo usuário ao chatterbot, um padrão de resposta, que será usado para montar uma sentença a ser retornada ao usuário, e um contexto opcional (DIAS, HENN, SILVA, 2007; WALLACE, 2012).

Além de realizar casamento de padrões (semelhante ao realizado pelo Eliza), a linguagem AIML introduz um conjunto de tags que a diferenciam do Eliza com relação a características como: possuir memória, e ser capaz de contextualizar e reavaliar sentenças digitadas pelo usuário (NEVES; BARROS, 2005).

Leonhardt (2005) apresenta as principais tags da linguagem no quadro a seguir.

Figura 5 – Quadro das principais tags da AIML

<aiml>	Inicia e termina um bloco de um arquivo programado em AIML
<category>	Identifica uma "Unidade de Conhecimento" na base de conhecimento, dentro desta tag que se encontra o padrão de entrada e a resposta para o usuário.
<pattern>	Identifica um padrão de mensagem que poderá ser digitado pelo usuário.
<template>	Contém a resposta para o usuário pelo padrão especificado.

Fonte: Leonhardt (2005).

Exemplificando a utilização destas tags, Leonhardt (2005) apresenta, na figura a seguir, como definir uma saudação:

Figura 6 – Saudação em AIML

```
<aiml>
  <category>
    <pattern> BOM DIA * </pattern>
    <template> LINDO DIA NAO ? </template>
  </category>
</aiml>
```

Fonte: Leonhardt (2005).

Primeiro há a abertura da tag <aiml>, indicando que o arquivo será uma base de conhecimento. Após, há a abertura da <category>, indicando o início de uma "Unidade de Conhecimento". Dentro das tags <pattern></pattern>, existe a saudação do usuário. E, dentro das tags <template></template>, a resposta que será dada pelo chatterbot quando o padrão de dentro do <pattern> for reconhecido. O asterisco (caracter *) representa um conjunto de caracteres (LEONHARDT, 2005).

A linguagem AIML não é apenas uma base de dados de perguntas e respostas, ela pode conter, dentro da tag <template>, a tag <srai> que, através de recursividade, consegue alcançar outras categorias possíveis (WALLACE, 2012).

Quanto a tag <srai>, Wallace (2012) fornece o seguinte exemplo:

Figura 6 – Quadro de saudação em AIML.

```
<aiml>
  <category>
    <pattern> DO YOU KNOW WHO * IS </pattern>
    <template><srai> WHO IS <star/></srai></template>
   </category>
</aiml>
```

Fonte: Wallace (2012).

Esse exemplo significa que, quando o padrão ‘*Do you know who * is ?*’ for encontrado como padrão de entrada do usuário, o chatterbot redirecionará sua pergunta para o padrão ‘*Who is **’.

Correa (2012) afirma quanto ao padrão da linguagem que é simples, contendo apenas palavras, números espaços e curingas (“_” e “*”).

Como essa é uma linguagem, baseada em XML, são necessários programas específicos para compreender o correto funcionamento do robô de conversação, a esses programas é dado o nome de interpretadores e são desenvolvidos nas mais diversas linguagens de programação. Como exemplo de interpretadores, pode-se citar o ProgramD, ProgramE, ProgramP e Program Y (CORREA, 2012).

“O AIML e o ALICE representam um ponto de partida para muitos outros projetos de chatterbots disponíveis hoje na internet. Para isso, basta que seja desenvolvida uma nova base de conhecimentos em AIML” (LEONHARD, 2005, p.28).

2.2.2 Modelo de Markov

Os chatterbots da segunda geração se utilizaram de técnicas da Inteligência Artificial como as Rede Neurais e o Modelo de Markov para sua implementação. Um dos chatterbots de grande reconhecimento que se utiliza do modelo de Markov é o MegaHAL, criado por Jason L. Hutchens (HUTCHENS; ALDER, 1998).

Esse sistema consiste em construir um modelo de linguagem baseado nas informações obtidas através das conversas com os usuários e, por meio desse modelo, ele escolhe a melhor resposta (JACOB JUNIOR, 2008).

Para definir o que é um Modelo Oculto de Markov, Russel e Norvig (2004, p. 533) afirmam “Um Modelo Oculto de Markov é um modelo probabilístico temporal no qual o estado do processo é descrito por uma única variável aleatória discreta. Os valores possíveis da variável são os estados possíveis do mundo”.

De início, a entrada do usuário é analisada em uma sequência de palavras e não-palavras, na qual uma palavra consiste em uma série de caracteres alfanuméricos e uma não-

palavra é uma série de outros caracteres (HUTCHENS, ALDER, 1998). A sequência resultante de palavras e não-palavras é utilizada para treinar dois Modelos de Markov de 4ª ordem, no qual, um dos modelos prevê qual símbolo ocorrerá depois de uma sequência de 4 símbolos, e o outro modelo prevê qual símbolo irá preceder a sequência (HUTCHES, ALDER, 1998).

Apesar de ter utilizado uma técnica inovadora, o MegaHAL não venceu o prêmio Loebner, ficando com a terceira colocação (JACOB JUNIOR, 2008).

2.2.3 Lógica Difusa

A lógica difusa, ramo da inteligência artificial que estuda a teoria dos conjuntos difusos, permite representar a pertinência a um conjunto, como a distribuição de possibilidades (RICH; KNIGHT, 1993).

A tecnologia, para a criação do chatterbot FRED, utiliza-se, além da técnica de aprendizado, baseado em frames, da lógica difusa para sua implementação. Essa é uma das principais tecnologias base dos chatterbots de sucesso em simular comportamento humano (LAUREANO, 1999).

O chaterbot FRED tem a habilidade de aprender, enquanto conversa com o usuário, pois, cada vez que encontra uma nova frase, ele responde com a frase adequada mais próxima em sua base de dados e, num estágio posterior, pode ensinar respostas para cada uma das frases encontradas (HUTCHENS, 1997).

Segundo Laureano (1999), o diferencial nesse chatterbot é a utilização da lógica difusa para montar um conjunto difuso de palavras chaves.

2.2.4 Raciocínio Baseado em Casos (RBC)

Para definir o que é o raciocínio baseado em casos, Fernandes (2005, apud MOURA, 2008) descreve que é um técnica da IA pela qual, dada uma situação atual para a qual não se obtenha uma resposta, busca-se a solução através da recuperação e adaptação de soluções passadas semelhantes, dentro do mesmo domínio do problema.

Muitos chatterbots se utilizam dessa técnica para melhorar sua interação com o usuário. A empresa Inference da Califórnia desenvolveu um chatterbot que possui um conjunto de casos passados em sua base de conhecimento, permitindo buscar nesse conjunto de dados uma solução e adaptando-a para o problema atual (LAUREANO, 2005).

Kraus e Fernandes (2007), na Univali, desenvolveram um chatterbot para a área imobiliária, utilizando-se dessa técnica, esse chatterbot busca os imóveis mais similares ao desejado pelo cliente, e os casos são representados pela característica de cada imóvel.

Sistemas desenvolvidos com esse paradigma possuem capacidade de aprendizado, pois um problema pode ser armazenado, após ter sido solucionado, e torna-se, então, disponível, caso futuramente haja um problema semelhante, a solução do problema anterior possa servir de resposta, aumentado o conhecimento existente no sistema (MELCHIORS, 1999 apud LEONHARDT, 2005).

2.3 PROBLEMÁTICA DOS CHATTERBOTS

Para o desenvolvimento de um chatterbot, é necessário estudar as problemáticas envolvidas em sua construção, tanto do ponto de vista computacional, quanto conversacional.

Chatterbots podem ser vistos como aplicações de processamento de linguagem natural e, dessa forma, enfrentam os mesmos problemas que esses sistemas enfrentam, como ambiguidade léxica e semântica (ALLEN; 1995, apud NEVES; BARROS; 2005).

Perante os problemas encontrados nos programas de processamento de linguagem natural, Specia e Nunes (2004) afirmam “Grande parte desses problemas está relacionada à ambiguidade inerente às línguas naturais, nos seus diversos níveis, como o morfológico, lexical, sintático, semântico, contextual e pragmático”.

Porém, como, no desenvolvimento de chatterbots, utiliza-se técnicas, além das utilizadas em sistemas de PLN, suas dificuldades não se limitam apenas aos mesmos problemas enfrentados por esses sistemas. Em sua construção, os chatterbots também enfrentam dificuldades específicas, como o controle do andamento global da conversação, o controle de sentenças repetidas e o tratamento de sentenças desconhecidas (NEVES; 2005, apud NEVES; BARROS; 2005).

As pesquisas, nas tecnologias de desenvolvimento dos chatterbots, estão cada vez mais avançadas, entretanto, o processamento de linguagem natural ainda esbarra em algumas barreiras. Conforme Leonhardt (2005), algumas das dificuldades encontradas são a ambiguidade de muitas sentenças e a falta de cobertura linguística e conceitual.

Segundo Primo e Coelho (2000), um dos problemas na tecnologia dos robôs de conversação é a pré-definição das respostas, todas as conversas sobre um determinado assunto precisam estar na base de conhecimento do sistema, pois se algum caso não for previsto, provavelmente, terá uma resposta padrão evasiva.

Além desses problemas, existem ainda aqueles relacionados à língua, pois, por maior que seja o vocabulário do robô, ainda, assim, existem aspectos como a singularidade de cada pessoa, os regionalismos, a variedade de significados, gírias e etc (MOURA, 2008).

As limitações dos chatterbots não se aplicam apenas à análise computacional do sistema, também existem barreiras relacionadas ao ponto de vista conversacional, e para explorar essas limitações, propõe-se um estudo da Teoria da Análise da Conversação (TAC). (NEVES; BARROS, 2005).

Devido a isso, percebe-se que muitos dos problemas estão relacionados ao aspecto do estudo da teoria da conversação, e não apenas do estudo computacional. Neves e Barros (2005) analisam três problemas recorrentes em chatterbots:

1) não levar em conta a estrutura global de uma conversação (abertura, desenvolvimento e fechamento);
2) tratar sentenças repetidas do usuário apenas com base na sua estrutura sintático-morfológica (não considerando “oi” e “olá” como repetição);
3) muitas sentenças são tratadas como desconhecidas, quando, na verdade, são turnos adjacentes. (por exemplo: quando o chatterbot pergunta “Você gosta de futebol ?” e o usuário responde “sim”, o programa, muitas vezes, não compreende a resposta, e nisso se perde a fluência do diálogo). Os turnos

adjacentes, também chamados de pares adjacentes, serão explicados mais adiante neste trabalho.

Além dos fatores problemáticos levantados até o momento, existe o questionamento do entendimento e da compreensão da conversa. Esses dois processos são fatores fundamentais na comunicação humana, pois ela não se resume em uma relação apenas de *inputs* e *outputs*, é preciso levar em consideração o *throughput*, ou seja, o que acontece entre o estímulo e a resposta (PRIMO; COELHO, 2000).

2.3.1 Teoria da Análise da Conversação

Analisando os problemas recorrentes em chatterbots, Neves e Barros (2005) chegaram a conclusão de que poderia ser utilizada a Teoria da Análise da Conversação para tratar alguns dos problemas encontrados.

Marcuschi (1986, apud NEVES; BARROS, 2005) propôs uma análise tanto no nível global quanto local de conversação. Destacou dois conceitos essenciais na organização local: o turno e os pares adjacentes e, na organização global, ele divide o diálogo em três grandes seções, sendo elas: a de abertura, a de desenvolvimento e a de fechamento.

2.3.1.1 Organização Global de Conversação

Segundo Canuto (2005) e Marcuschi (2003), uma conversa considerada normal acontece em três fases: a abertura, o desenvolvimento e o fechamento. Na fase de abertura, é quando se realiza o contato inicial: as interações, como saudações, cumprimentos e apresentações. Na fase de desenvolvimento, iniciam-se os tópicos da conversa, e os assuntos são desenvolvidos, enfatizando que uma conversa não precisa ter apenas um tópico, ela pode ter vários tópicos. Na última fase, a de fechamento, é quando ocorre a despedida.

2.3.1.2 Organização Local de Conversação

Levinson (1983, apud MARCUSCHI, 2003, p.14) afirma: “A conversação é a primeira das formas de linguagem a que estamos expostos e, provavelmente, a única da qual nunca abdicamos pela vida afora”.

Para manter uma conversação, existem alguns pontos básicos em comum. Marcuschi (2003) afirma que, para construir e manter uma conversação, duas pessoas devem partilhar um mínimo de conhecimentos comuns, entre eles estão a aptidão linguística, o envolvimento cultural e o domínio de situações sociais. Não se limitando a apenas esses fatores, também, é necessário aos participantes prestarem atenção aos fatos paralinguísticos, como os gestos, os olhares, os movimentos do corpo e outros. Porém, apenas, o domínio da língua não garante o sucesso do objetivo, é necessário haver aptidões cognitivas para tal.

Para a organização local da conversação, Marcuschi (2003) apontou a existência de uma série de características organizacionais, entre elas estão a organização de turno a turno e a organização de sequências.

Numa conversação, a unidade é o turno, em um certo momento, um falante está no turno de fala, posteriormente, um ouvinte toma o turno pra sí e torna-se o falante atual e, nisso, tem-se uma troca de turnos (CORREA; 2012). A conclusão de um turno pode acontecer a qualquer momento, sendo difícil definir o que causa a mudança de turno entre os participantes da conversa. Marcuschi (2003) aponta algumas regras e mecanismos para tratar este problema, como o “Fala um por vez”, “Quem tem a palavra e quando”, “Fala simultâneas e sobreposições”, “Pausas, silêncios e hesitações” e “Reparações e Correções”.

Dentro da organização de sequências, está o conceito de pares adjacentes. Marcuschi (2003) os aponta como sendo:

Pergunta – Resposta;

Ordem – Execução;

Convite – Aceitação/Recusa;

Cumprimento – Cumprimento;

Xingamento – Defesa/Revide;

Acusação – Defesa/Justificativa;

Pedido de Desculpa – Perdão;

Schegloff (2002, apud NEVES; BARROS, 2005, p.1036) informa: “A intenção embutida na segunda parte de um par adjacente pode ser prevista e varia de acordo com o contexto sociocultural em que se dá a conversação”. Segundo Neves e Barros (2005), o tratamento da intenção é o estudo, dentro da linguística, que atribui significado a fala através da interpretação da intenção do falante, e a noção de intenção é considerada pela Teoria da Análise da Conversação no processo da análise de diálogos.

2.3.2 Ambiguidade Léxica de Sentido

Outro problema encontrado na construção de chatterbots, como relatado neste trabalho anteriormente, é a questão da ambiguidade léxica.

A ambiguidade léxica acontece quando uma palavra tem mais de um possível significado, para exemplificar, pode-se citar a palavra “bala”, que pode ser utilizada no sentido de um doce ou de uma munição (COPPIN, 2012).

Segundo Vieira e Lima (2012), a semântica lexical considera as propriedades referentes a cada uma das palavras da língua, e um dos primeiros problemas encontrados é o fato de uma palavra apresentar múltiplos significados.

O problema da ambiguidade lexical veio da necessidade de escolher por um dos possíveis significados de uma palavra quando da sua interpretação, e sua causa fundamental é a existência de relações semânticas interlexicais como a polissemia e a homonímia (SPECIA; NUNES, 2004).

Polissemia, segundo (PRIBERAM, 2012), é o “Conjunto dos vários sentidos de uma palavra ou locução”. Na polissemia, uma palavra tem dois ou mais significados, mas relacionados entre si, sendo que, normalmente, apenas um dos significados se ajusta a um determinado contexto, já, na homonímia, duas ou mais palavras com significados totalmente diferentes, sem traços comuns são idênticas quanto ao som e/ou grafia (SPECIA; NUNES, 2004).

A polissemia e a homonímia são assuntos muito discutidos pelos gramáticos da língua portuguesa. Quanto a polissemia, Lima (2005, apud LIMA, 2012) apresenta como “a multiplicidade de sentidos imanente em toda palavra, que possui estrita dependência do contexto e que tem como resultado a sinonímia.”, já, Bechara (2004, apud LIMA, 2012) apresenta a polissemia sendo: “O fato de haver uma só forma (significante) com mais de um significado unitário pertencentes a campos semânticos diferentes. [...] cada um desses significados é preciso e determinado”.

A homonímia, para Lima (2005, apud LIMA, 2012) é um “fator de perturbação da boa escolha das palavras”, para Júnior (1985, apud LIMA, 2012), ela pode ser considera como “A propriedade de duas ou mais formas, inteiramente distintas pela significação ou função, terem a mesma estrutura fonológica: os mesmos fonemas dispostos na mesma ordem e subordinado ao mesmo tipo de acentuação”.

Como exemplo de homonímia, pode-se citar a palavra “cheque”, que significa ordem de pagamento, e “xeque” que é uma jogada do xadrez. As duas palavras pronunciam-se iguais, porém sua escrita e seu significado são distintos.

No estudo dos chatterbots, neste trabalho, será dada uma ênfase às palavras polissêmicas.

Saramento (2006) afirma que a existência de palavras polissêmicas introduz um fator de ambiguidade que dificulta a sua análise automática, e cita o exemplo da palavra “laranja”, que pode ser encontrada tanto no conjunto das cores, quanto no conjunto das frutas.

Divide-se o problema da ambiguidade lexical em ambiguidade categorial ou ambiguidade de sentido (ULLMAN, 1964, apud SPECIA; NUNES, 2004).

A ambiguidade categorial acontece quando os significados da palavra polissêmica são de classes gramaticais diferentes, como ocorre na tradução do inglês da palavra “field”, que pode ser traduzida para “campo”, sendo assim um substantivo, ou “interceptar” sendo um verbo; a ambiguidade de sentido, porém, trata quando as duas ou mais opções de significado para uma palavra polissêmica são da mesma classe gramatical (SPECIA; NUNES, 2004).

O problema de determinar qual dos sentidos será utilizado num determinado contexto é conhecido como “Desambiguação de Sentidos” e representa uma área de estudo muito ativa (SARAMENTO, 2006).

A ambiguidade categorial pode ser resolvida através da análise das características sintáticas da palavra, com procedimentos como a análise sintática ou a etiquetação gramatical, obtendo resultados muito satisfatórios nas pesquisas, contudo, a ambiguidade de sentidos

exige a análise da semântica e do estudo de tais palavras, sendo o foco da maioria dos trabalhos voltados para o tratamento da ambiguidade lexical (SPECIA; NUNES, 2004).

Coppin (2012) afirma que, na ambiguidade léxica, um analisador pode determinar qual parte do discurso é pretendida, porém, em alguns, casos é necessária desambiguação semântica para determinar qual sentido é pretendido.

O processo de um sistema de processamento de linguagem natural em determinar qual o sentido é pretendido por uma frase ambígua é conhecido como desambiguação, e ela pode ser feita de diferentes formas (COPPIN, 2012).

Conforme Specia e Nunes (2004), para realizar a desambiguação de forma automática, é necessário haver um módulo de Desambiguação do Sentido das Palavras, também chamado de Desambiguação Léxica de Sentido (DLS), aos processos de interpretação e/ou geração da língua; para construir esse módulo, são necessárias serem analisadas algumas questões, como, por exemplo: palavras a desambiguar e quais os possíveis sentidos dessas palavras, qual método será adotado para a desambiguação e de como será avaliado o módulo.

Saramento (2006) partiu da ideia de que uma palavra polissêmica, para um determinado sentido, co-ocorre junto com um conjunto de palavras típicas desse sentido, dessa forma utiliza-se de técnicas com algoritmos de agrupamento com o objetivo de determinar os possíveis sentidos pretendidos de uma palavra, entre estes algoritmos estão o PAM e o CLARA, que obtiveram melhores resultados que os demais.

Coppin (2012) explica que um dos métodos mais eficaz para realizar a desambiguação é utilizando probabilidade, podendo ser a priori ou condicional. A probabilidade a priori serve para dizer ao sistema que a palavra “bala” quase sempre se refere a doce, e a probabilidade condicional diria que ao usar a palavra bala por crianças, ela se referiria a doce, porém ao utilizar a mesma palavra por um colecionador de armas, seria mais provável tratar-se de munição. Coppin ainda afirma que o contexto também é extremamente importante na tarefa de desambiguação.

Vieira e Lima (2001) apresentam a ideia do “Léxico” ou “Dicionário” para o tratamento da ambiguidade, o definindo como uma estrutura fundamental para a maioria dos sistemas de PLN sendo uma estrutura de dados com os itens lexicais e as informações correspondentes a estes itens. Cada item pode ser uma palavra, “mel”, por exemplo, ou palavras que juntas apresentam um significado específico, como “lua de mel”. Vieira e Lima (2012), ainda, afirmam que “O léxico irá representar, através das múltiplas descrições que podem estar associadas a uma entrada, os múltiplos sentidos de cada palavra ou item lexical”.

Duas formas de reunir os itens em um léxico para tratar a ambiguidade são a “estrutura de formas”, que se utilizam de conter no léxico as unidades (palavras ou unidades maiores que palavras) por extenso, como, por exemplo: a palavra “casa” se referindo a “ substantivo feminino singular normal” e “casa” se referindo a “verbo singular 3ª pessoa presente indicativo 1ª conjugação”. A outra forma é a “estrutura de bases”, que consiste em guardar no léxico unidades menores que concentram a capacidade de identificar um determinado item, como, por exemplo “cas” para “casa” e “preven” para “prevenção” e “prevenir” (VIEIRA; LIMA, 2012)

Para trabalhar com os sentidos, é interessante haver uma organização em classe de objetos, de acordo com como usualmente classificamos as coisas do mundo. Tais classificações são chamadas de taxonomias ou ontologias (VIERIA; LIMA, 2012).

2.4 TRATAMENTO DA DESAMBIGUAÇÃO LÉXICA

Na sessão anterior deste trabalho, foi apresentada a problemática da ambiguidade léxica, e o porquê dela ocorrer, na sessão atual serão demonstradas algumas técnicas utilizadas para tratar essa ambiguidade.

O tratamento da ambiguidade léxica é um estudo amplo na área de processamento de linguagem natural, que engloba trabalhos realizados para áreas como a Tradução Automática (SPECIA; NUNES, 2004), a Mineração de Textos e a Recuperação da Informação (SAYÃO, 2007).

A DLS é descrita como um problema de "AI-complete" (MALLERY, 1988, apud NAVIGLI, 2009) o que significa que sua dificuldade é equivalente a resolver os problemas centrais da inteligência artificial, como, por exemplo, o teste de Turing (NAVIGLI, 2009).

Isso acontece, primeiramente pelo fato de levar a diferentes formalizações de questões fundamentais como a representação dos sentidos de cada palavra, a granularidade dos catálogos de sentidos, a utilização ou não de um domínio específico, o conjunto de palavras a desambiguar, etc. Segundo que depende fortemente de conhecimento, a estrutura de um sistema DLS pode ser resumida em: dado um conjunto de palavras, uma técnica utilizando alguma fonte de conhecimento é utilizada para associar os sentidos mais apropriados as palavras do contexto (NAVIGLI, 2009).

Segundo Ide e Véronis (1998) e Specia e Nunes (2004), na criação de modelos para a tarefa de desambiguação léxica de sentido, os seguintes passos são considerados:

1) determinar o conjunto de palavras a serem desambiguadas: Todas as palavras, apenas as palavras de uma determinada classe gramatical ou um subconjunto de palavras, etc;
2) definir todos os sentidos possíveis de cada palavra: Na DLS monolíngue são os significados da palavra. Criar um mecanismo para atribuir a cada ocorrência da palavra o sentido mais apropriado;
3) avaliar este mecanismo estabelecido.

Determinar o conjunto das palavras que precisarão passar pelo tratamento da desambiguação depende da aplicação do mecanismo de DLS. Alguns trabalhos realizam a desambiguação de todas as palavras de conteúdo da língua. Esses trabalhos são utilizados na categorização de textos em que o sentido das palavras pode auxiliar a identificar a área; em outros trabalhos, identificam-se apenas as palavras com que a ambiguidade pode ocasionar algum problema, como nos sistemas de Tradução Automática. Em geral, a desambiguação de todas as palavras é mais robusta, enquanto que a de subconjuntos é mais precisa. (SPECIA; NUNES, 2004).

Conforme Kilgarrif (1997, apud SPECIA; NUNES, 2004), não é possível definir os sentidos das palavras com base em quaisquer recursos lexicais, ele acredita que tais sentidos devem ser abstrações criadas especificamente para cada tarefa de DLS a partir de citações de córpus e ainda apresenta duas formas para determinar o conjunto de sentidos de uma palavra: a frequência e a imprevisibilidade.

Para melhor entendimento dessas duas formas: “Deve-se considerar um sentido como um dos possíveis sentidos da palavra se sua frequência no córpus for alta e se tal sentido não puder ser previsto ou derivado a partir do sentido básico” (KILGARRIF, 1997, apud SPECIA; NUNES, 2004, p.6).

O conceito de córpus será apresentado mais a frente, neste trabalho, ao distinguir os métodos de DLS em: método baseado em conhecimento, método baseado em córpus, método híbrido e método baseado em IA (OLIVEIRA NETO, 2004; IDE; VERONIS, 2001).

Segundo Navigli (2009) fontes de conhecimento são componentes fundamentais de um módulo DLS, elas provêm de dados que são essenciais para associar os sentidos às suas palavras, podendo variar entre “corpora” de textos, tesauros, dicionários, glossários, ontologias e muitos outros. Neste trabalho, serão apresentadas apenas as fontes de

conhecimento que serão utilizadas no desenvolvimento do chatterbot, se limitando a "*Bag-of-Words*", "*Collocations*" e "Associação Semântica de Palavras (*Semantic Word Associatcion*)" que serão citadas mais a frente.

"Uma questão importante na definição de um trabalho de DLS é a definição dos tipos de conhecimento que serão empregados e das fontes de informação que podem ser utilizadas para prover esses conhecimentos" (SPECIA; NUNES, 2004, p. 11).

Segundo Specia e Nunes (2004), a desambiguação pode ser realizada com base em dois grupos de conhecimento: o conhecimento do contexto da palavra-alvo (palavra a ser desambiguada), que contém informações do próprio texto do qual a palavra faz parte e informações extralinguísticas sobre o texto e o conhecimento externo, que inclui tipos de conhecimento especificados manualmente ou advindos de recursos lexicais.

2.4.1 Fontes de Conhecimento

A seguir, serão apresentadas apenas algumas das fontes de conhecimento estudadas, pois tentará se focar na utilização futura dessas abordagens para o desenvolvimento do chatterbot. Os tipos de conhecimento que serão mostrados, a seguir, são: "*Bag-of-Words*", "*Collocations*", "Associação Semântica de Palavras (*Semantic Word Association*)" e *Stoplist*.

2.4.1.1 Bag-of-Words

Hahn (2008) afirma que um método comum de extrair características de um texto é através do modelo de *bag-of-words*. Este modelo consiste em extrair de um texto as características dele contando o número de ocorrências de cada palavra.

Segundo Sayão (2007), uma das abordagens para a estruturação de documentos é a *bag-of-words*, em que cada documento é representado por um vetor dos termos existentes

nele. Cada coluna do vetor representa um termo, e em cada célula existe o peso do termo, indicando a relevância para o documento.

Metzeler Jr (2007) afirma que o que todos os modelos de *bag-of-words* tem em comum é que não levam em consideração a ordem em que os termos aparecem no documento.

Outra definição, para essa abordagem, é feita por Specia e Nunes (2004), que a descrevem como:

> Conjunto de palavras que circundam a palavra a ser desambiguada em uma janela que pode variar desde uma quantidade pré-determinada de palavras na sentença ou no texto no qual a palavra está localizada à sentença inteira ou algumas sentenças do texto. Normalmente essa janela considera a palavra ambígua como centro e uma igual quantidade máxima de palavras em ambos os lados, sem considerar a sua ordem e independentemente das suas características.

Specia e Nunes (2004) ainda afirmam que em geral as palavras são lematizadas[2] e algumas são excluídas através de uma lista de *stop-words*[3], fazendo dessa forma a captura do tópico geral do texto mais facilmente; como exemplo pode-se descrever um texto com as palavras próximas "*bank*", "*loan*", e "*payment*" levando a crer que a palavra "*interest*" esteja atribuída ao significado de "juros".

Um dos problemas enfrentados na utilização desta abordagem é quando ocorre a existência de mais de um documento, gerando uma matriz termo-documento onde cada linha refere-se a um documento, esta matriz acaba tomando uma dimensão muito grande, gerando uma matriz esparsa, pois todos os termos tem que estar presentes (SAYÃO, 2007).

Abaixo segue uma imagem representando uma matriz termo-documento retirada de Sayão (2007).

Figura 7 – Exemplo de uma matriz termo-documento

	$termo_1$	$termo_2$	$termo_3$	$termo_4$	.....	$termo_t$
doc_1	$peso_{11}$	$peso_{12}$	$peso_{13}$	$peso_{14}$	.....	$peso_{1t}$
doc_2	$peso_{21}$	$peso_{22}$	$peso_{23}$	$peso_{24}$	.....	$peso_{2t}$
.....	.....	.....	.....	.....	.....	.....
doc_n	$peso_{n1}$	$peso_{n2}$	$peso_{n3}$	$peso_{n4}$	.....	$peso_{nt}$

Fonte: Sayão (2007, p. 42).

[2] Lematização: Identificar o radical da palavra, excluindo o gênero, o plural, o tempo verbal, e etc.

[3] *Stop-words*: Palavras irrelevantes que formam a *Stoplist*, explicada mais adiante neste trabalho.

2.4.1.2 Collocations

O termo "*Collocation*", popularizado por J. R. Firth em 1951 no artigo de título "*Modes of Meaning*", vem sendo utilizado em muitos trabalhos de DLS. O autor afirma que uma "*collocation*" não é uma simples co-ocorrência de palavras, mas uma co-ocorrência habitual ou usual (IDES; VERONIS, 1998).

Barry-Rogghe (1973, apud IDES; VERONIS, 1998) define uma "collocation" como uma associação sintagmática entre itens lexicais em que a probabilidade do item x co-ocorrer com os itens a,b,c.. é maior do que a simples possibilidade.

Uma definição semelhante é a dada por Kilgarrif (1997, apud SPECIA; NUNES, 2004) em que afirma que uma "collocation" pode ser definida como um grupo de duas ou mais palavras encontradas próximas, com uma frequência significativamente maior do que apenas as palavras individualmente; elas ainda podem ou não serem vizinhas imediatas e o significado do todo pode ou não ser determinado pelo significado das partes.

Agirre e Martinez (2001, p.2) mostram um bom exemplo ao dizerem que dos sentidos do substantivo "match" do inglês, apenas um se encaixa na frase "football match".

Segundo Laffe (1998), a *collocation* substitui com muitas vantagens o uso de uma enciclopédia para resolver os problemas de ambiguidades lexicais.

2.4.1.3 Associação Semântica de Palavras (*Semantic Word Association*)

Agirre e Martinez (2001) afirmam que as "*collocations*" e a associação semântica de palavras são as fontes de conhecimento mais importantes para uma aplicação de desambiguação léxica de sentido.

Diante dessaa observação, é interessante apresentar o que é a associação semântica de palavras. Specia e Nunes (2004) e Agirre e Martinez (2001) falam sobre essa fonte de conhecimento e argumentam que podem ser separadas em quatro tipos:

1) organização de taxonomias;

2) situação;

3) tópico;

4) relação núcleo-argumentos (*Argument-head relation*).

Nessas associações, se dadas como uma relação sentido-palavra, tem-se um indicador muito forte do sentido pretendido. Um exemplo seria na frase "*the chair and the table were missing*", no qual a palavra "*chair*", podendo ter diversos significados, pelo fato de compartilhar uma classe na taxonomia com a palavra "*table*" pode ser utilizada no sentido de mobília. Outro exemplo seria, quando falando sobre "*baseball*", é mais provável que a palavra "*bat*" esteja relacionada a "bastão" do que a "morcego" (AGIRRE; MARTINEZ, 2001).

2.4.1.4 Stoplist

Outra fonte de conhecimento citada por Navigli (2008), em seu trabalho, é a *Stoplist*. Barion e Lago (2008), em seu artigo sobre mineração de textos, descrevem essa fonte de conhecimento como uma lista na linguagem que se está trabalhando, contendo palavras irrelevantes (*stopwords*). Esse processo Stoplist consiste em remover um conjunto de palavras que são consideradas irrelevantes para a extração da informação, essas palavras (*stopwords*) geralmente são preposições, artigos, conjunções, alguns verbos, nomes, adjetivos e advérbios.

2.4.1.5 Outras Fontes de Conhecimento

Muitas outras fontes de conhecimento, além das apresentadas nesta monografia, são citadas nos trabalhos de desambiguação léxica de sentido. Abaixo, será apresentada uma lista dessas fontes, de acordo com Navigli (2009), Agirre e Martinez (2001), Specia e Nunes (2004):

- tesauros;
- ontologias;

- corpora;
- etiquetas gramaticais;
- informações sintáticas;
- papéis semânticos;
- preferências de seleção;
- domínio;
- frequência de sentidos;
- pragmática.

Essas fontes, porém, apesar de serem muito importantes no estudo da desambiguação léxica de sentido, não serão descritas mais profundamente, neste trabalho, pelo fato de não fazerem parte das fontes escolhidas para o desenvolvimento do chatterbot.

2.4.2 Métodos de Desambiguação Léxica de Sentido

Nesta sessão do trabalho, serão apresentados os métodos de desambiguação léxica de sentido.

Segundo Oliveira Neto (2004), os trabalhos de DLS podem seguir diferentes métodos da PLN, e cita três: O método baseado em conhecimento linguístico manualmente especificado, o método baseado em extração de conhecimento a partir de corpus de exemplos, e o método híbrido, que utiliza ambos os métodos anteriores.

Já, para Ide e Veronis (1998), além de separar também em método baseado em corpus e método baseado em conhecimento, eles ainda apresentam outra categoria, a do método baseado em técnicas da inteligência artificial.

A seguir, serão apresentados os métodos de desambiguação léxica de sentido.

2.4.2.1 Método Baseado em Inteligência Artificial

Na década de 1960, surgiram as técnicas de Inteligência Artificial (IA) e, com isso, começou-se a tratar os problemas de PLN através desta abordagem, resultando em trabalhos com a pretensão de realizar o entendimento completo da língua, utilizando-se de conhecimentos detalhados de sintaxe e semântica para realizar as tarefas desejadas (IDE; VERONIS, 1998).

Logo após o desenvolvimento das redes semânticas, no final dos anos 1950, essas estruturas já começaram a ser aplicadas para a representação do sentido de palavras (IDE; VERONIS, 1998).

Ide e Veronis (1998), ainda, separam o método baseado em Inteligência Artificial em dois: os Métodos simbólicos e os Métodos Conexionistas, entretanto, Specia e Nunes (2004) agrupam essas técnicas juntas com as técnicas do método baseado em conhecimento.

Entre os métodos simbólicos, pode-se citar o trabalho de Quillian (1961, 1962a,1962b, 1967, 1968, 1969, apud IDE; VERONIS, 1998), em que o autor constrói uma rede que inclui ligações entre palavras e conceitos, e as ligações são rotuladas com relações semânticas ou simplesmente associações entre palavras.

Outros trabalhos também na área de métodos simbólicos são os descritos por Specia e Nunes (2004) e Ide e Veronis (1998), nos quais se utilizam técnicas como redes semânticas em conjunto com frames, semântica de preferência, características sintáticas e semânticas aliadas à semântica de preferência e outras.

Os métodos conexionistas utilizam a técnica de "*semantic priming*", fundamentada em trabalhos da Ciência Cognitiva, citada por meio de redes neurais artificiais que se mostram muito apropriadas para o tratamento das tarefas de DLS (SPECIA; NUNES, 2004).

Specia e Nunes (2004) afirmam "Os modelos de '*semantic priming*' são geralmente baseados em conceitos de ativação propagada. Neles, a representação mental de conceitos é uma rede, em que conceitos semanticamente relacionados estão próximos uns dos outros".

Segundo Squire e Kandel (2003, apud Salles et al. 2007):

> Priming é um tipo de memória implícita (não declarativa) referente aos efeitos facilitadores de eventos antecedentes (primes) sobre o desempenho subsequente (respostas aos alvos), ou seja, um aperfeiçoamento da capacidade de detectar ou identificar palavras, objetos ou figuras após uma experiência recente com eles.

2.4.2.2 Método Baseado em Conhecimento

Os trabalhos desenvolvidos nos anos 70 e 80 se utilizando de métodos da I.A demonstraram-se muito eficazes para casos em que o domínio fosse bem limitado. Porém existia a dificuldade de elaborar e rotular as grandes quantidades de conhecimento existentes para o domínio especificado.

Com o advento dos recursos lexicais como dicionários, tesauros e corpora, esses recursos passaram a ser amplamente utilizados em conjunto com técnicas para extração de conhecimento (IDES; VERONIS, 1998).

Segundo Navigli (2009), o objetivo do método, baseado em conhecimento, que ele também denomina como método baseado em dicionário, é explorar o conhecimento em recursos, como dicionários, tesauros, ontologias, *collocations* e etc. para deduzir o sentido da palavra ambígua no contexto relacionado.

"Os dicionários eletrônicos constituem fontes de informação para a desambiguação extraídas automaticamente a partir de recursos não eletrônicos já existentes, criados com outras finalidades, normalmente para uso por seres humanos" (SPECIA; NUNES, 2004).

Como exemplo destes dicionários eletrônicos, pode-se citar os MRD (*Machine-Readable Dictionaries*), que se tornaram populares fontes de conhecimento para o processamento de linguagem natural, sendo que, na década de 1980, os trabalhos em DLS se focaram em extrair automaticamente conhecimentos lexicais e semânticos dessas bases (IDE; VERONIS, 1998).

A ideia principal dessa abordagem, utilizando dicionários, como afirma Preiss (2006) é a de que palavras relacionadas entre si, também terão palavras em comum em suas definições, e funciona escolhendo os sentidos que maximizam a sobreposição de palavras em suas definições.

Lesk (1986, apud PREISS, 2006) afirma que escolher os sentidos que maximizam a sobreposição de palavras em suas definições é uma maneira não-supervisionada de explorar um MRD, pois não há utilização de rotulação e etiquetação das palavras.

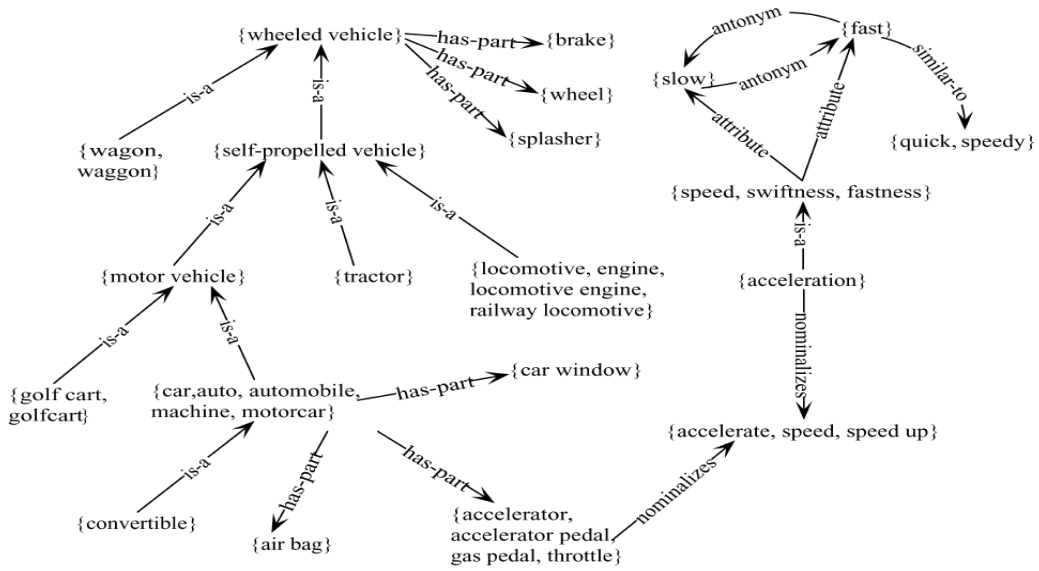
Segundo Ide e Veronis (1998), no meio dos anos 1980, começou a se construir a mão bases de conhecimento de larga escala como a WordNet.

A WordNet é um léxico computacional na língua inglesa, baseado em princípios psicolinguísticos, criado e mantido pela Universidade de Princeton, que codifica os conceitos em conjuntos de sinônimos chamados de synsets (MILLER et al. 1990; FELLBAUM, 1998, apud NAVIGLI, 2009).

Quanto aos léxicos computacionais, Specia e Nunes (2004, p.22) afirmam que "[...] são recursos lexicais criados (em geral, manualmente), especificamente, para o tratamento computacional". Esses recursos ainda visam a permitir a representação, assim como a manipulação de diversos tipos de informação sobre cada item lexical (SPECIA; NUNES, 2004).

A seguir, é apresentada uma imagem de uma rede semântica da WordNet retirada de Navigli (2009).

Figura 8 – Imagem da rede semântica da WordNet



Fonte: Navigli (2009)

Outro recurso utilizado é o tesauro (thesauri). Segundo Navilgi (2009), os tesauros provêm informação sobre o relacionamento entre as palavras, como sinonímia e antonímia, e o tesauro mais utilizado no campo de DLS é o "*Rogets International Thesaurus*".

Os autores que utilizam este recurso afirmam que, através das categorias de uma palavra em um contexto, deduz-se qual a categoria semântica do contexto como um todo, e essa categoria determina os sentidos das palavras (SPECIA; NUNES, 2004).

2.4.2.3 Método Baseado em Córpus

Segundo Ide e Veronis (1998), a corpora é utilizada, desde a metade do século XX, sendo que, desde o fim do século XIX, a análise manual da corpora possibilitou o estudo de palavras, grafemas e a extração de listas de palavras e *collocations* para os estudos de aquisição de linguagem e ensino de línguas.

"[...] com os avanços na área de Aprendizado de Máquina (AM), tem crescido no PLN a utilização de métodos que permitem extrair conhecimento automaticamente a partir de corpus, visando minimizar o problema do gargalo da aquisição de conhecimento" (SPECIA; NUNES, 2004, p. 24).

Conforme Navigli (2009), o corpus é uma coleção de textos usados para aprender modelos de linguagem, sendo que ele pode ser anotado (etiquetado com os sentidos) ou não anotado, ambos os tipos de recursos são utilizados na DLS, e são muito úteis nas abordagens supervisionadas e não-supervisionadas, respectivamente.

Specia e Nunes (2004) afirmam quanto ao corpus que ele fornece um conjunto de exemplos que, quando submetidos a algoritmos de Aprendizado de Máquina, permitem a criação de modelos capazes de descrever esses exemplos e de prever o comportamento de novos exemplos, e muitos trabalhos realizam a desambiguação de sentido automaticamente a partir de um corpus.

Em um corpus anotado, no qual existem as etiquetações de sentido nos exemplos, utiliza-se de técnicas de aprendizado de máquina para extrair o sentido do rótulo, fazendo parte da abordagem de métodos supervisionados. Por outro lado, quando o corpus não é anotado, são aplicados métodos não-supervisionados (NAVIGLI, 2009).

Quanto ao aprendizado de máquina, Coppin (2012) afirma que o aprendizado supervisionado é aquele em que as redes neurais apresentam dados de treinamento pré-classificados, enquanto que o aprendizado não-supervisionado ocorre sem qualquer intervenção humana.

Specia e Nunes (2004) apresentam as vantagens e desvantagens desse método, que serão listadas abaixo:

Vantagens

1) não é necessário codificar todo o conhecimento manualmente;

2) utilizam-se algoritmos tradicionais de aprendizado de máquina;

3) os modelos criados são facilmente gerenciáveis;
4) os modelos gerados podem expressar algum conhecimento novo.

E como Desvantagens:

1) o corpus para a criação do modelo precisa ser representativo para o domínio do problema;
2) nos trabalhos supervisionados a etiquetação normalmente é feita manualmente;
3) não há certeza que os resultados serão adequados, problema decorrente do processo de aprendizado.

2.4.2.4 Método Híbrido

O método híbrido se utiliza da junção de características do método baseado em conhecimento e do método baseado em corpus (SPECIA; NUNES, 2004).

Oliveira Neto (2004) cita Zinovjeva (2000) como trabalho híbrido de DLS para a TA (Tradução Automática), em que traduz palavras ambíguas do inglês para o sueco em textos irrestritos de qualquer gênero ou domínio.

Nessa sessão, foram vistas as classificações dos métodos de desambiguação léxica de sentido, e é interessante resumir o que foi apresentado anteriormente, através de um quadro, para melhor entendimento e distinção dos conceitos de cada método.

Figura 9 – Quadro do Resumo dos métodos e de suas técnicas utilizadas

Método	**Técnicas Utilizadas**
Baseado em Inteligência Artificial	Métodos como redes neurais e ativação propagada.
Baseado em Conhecimento	Recursos léxicos como dicionários, tesauros, ontologias, etc.
Baseado em Córpus	Exemplos da língua para extrair conhecimento, podendo estes corpus serem anotados ou não-anotados.
Híbrido	Utiliza da junção de características dos métodos baseado em conhecimento e do método baseado em corpus.

Fonte: Elaboração do autor (2012).

3 MÉTODO

Nesse capítulo, será abordada a metodologia utilizada para a elaboração deste projeto, a caracterização do tipo de pesquisa, as etapas metodológicas, a proposta de solução para chegar ao objetivo estabelecido e as delimitações do projeto.

3.1 CARACTERIZAÇÃO DO TIPO DE PESQUISA

Primeiramente, é importante a definição do que é método. Segundo Galliano (1986, p. 6), "Método é um conjunto de etapas, ordenadamente dispostas, a serem vencidas na investigação da verdade, no estudo de uma ciência ou para alcançar determinado fim". Também, falando sobre o método, Cervo e Bervian (1996, p. 46) afirmam que "Métodos são técnicas suficientemente gerais para se tornarem procedimentos comuns a uma área das ciências ou a todas as ciências".

Quanto à pesquisa, ela pode ser distinguida em pesquisa pura, também chamada de pesquisa básica, em que o pesquisador tem como objetivo o saber, ou seja, obter o conhecimento, e a pesquisa aplicada, em que o investigador é motivado pela necessidade de colaborar para fins práticos, mais ou menos imediatos, a fim de buscar respostas e soluções para problemas concretos (CERVO; BERVIAN, 1996). A pesquisa prática, segundo Demo (2004), é "destinada a intervir diretamente, na realidade, a teorizar práticas, a produzir alternativas concretas e a comprometer-se com soluções". Este trabalho, portanto, pode ser inserido na classificação de pesquisa aplicada.

Santos (2002) afirma que as pesquisas podem ser caracterizadas segundo os objetivos, as fontes de dados e os procedimentos de coleta de dados.

Segundo os objetivos, este trabalho se enquadrou como uma pesquisa exploratória. Conforme Santos (2002), a pesquisa exploratória é o primeiro contato do pesquisador com o tema, pretendendo criar maior familiaridade com o mesmo, e isso se dá através da prospecção de materiais que informem ao pesquisador a importância do problema, o que existe sobre o assunto, em que estado se encontram as informações referentes ao

assunto e revelar novas fontes de informação, por isso ela é feita geralmente através de levantamento bibliográfico, entrevistas com profissionais da área e etc.

Por se tratar de um assunto muito específico, esta pesquisa foi realizada consultando principalmente materiais acadêmicos, tais como, relatórios, artigos e teses e, além desses, também foram utilizados livros da área para se ter um conhecimento mais fortemente embasado. Santos (2002, p. 27) afirma que bibliografia "É o conjunto de materiais escritos/gravados, mecânica ou eletronicamente, que contém informações já elaboradas e publicadas por outros autores", portanto pode-se caracterizar, de acordo com as fontes e coletas de dados, esta pesquisa como sendo bibliográfica.

3.2 ETAPAS METODOLÓGICAS

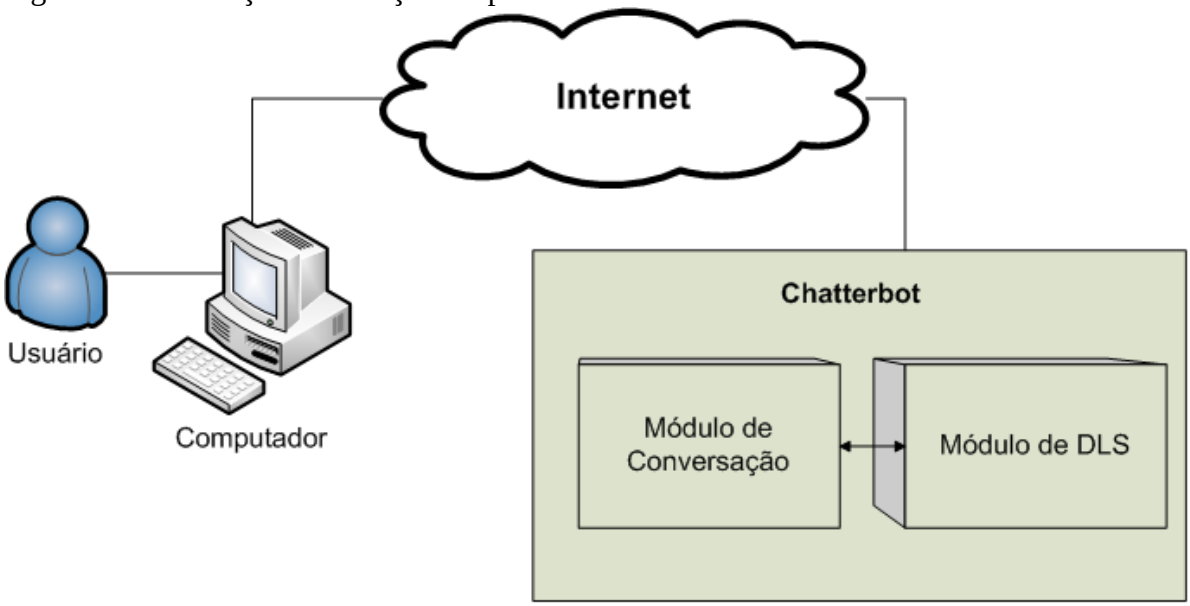
Este projeto possui as seguintes etapas descritas abaixo para alcançar o sucesso no objetivo requerido:

a) definir a base de dados que conterá o domínio de conhecimento, ou seja, que especificará qual será o assunto sobre o qual o chatterbot conversará com os usuários;
b) definir as palavras ambíguas dentro do domínio de conhecimento do chatterbot;
c) definir a utilização do método mais adequado para a implementação do módulo de conversação;
d) implementar o módulo de conversação;
e) definir quais os métodos e fontes de conhecimento que serão utilizados para o módulo de Desambiguação Léxica de Sentido (DLS);
f) implementar o módulo de DLS.

3.3 SOLUÇÃO PROPOSTA

Como solução proposta para este trabalho, pretende-se separar o sistema em dois módulos: o módulo de conversação, em que se inclui a implementação de alguma das técnicas utilizadas para a conversa do chatterbot com o usuário, como as descritas na seção 2.2 deste trabalho, e o módulo de desambiguação léxica de sentido (DLS), que conterá a implementação das técnicas e fontes de conhecimento que serão estabelecidos para efetuar a desambiguação de palavras.

Figura 10 – Ilustração da Solução Proposta



Fonte: Elaboração do Autor, 2012

3.3.1 Módulo de Conversação

O Módulo de Conversação será responsável por interagir diretamente com o usuário. Este módulo terá como função receber a entrada digitada pelo usuário e retornar uma resposta em linguagem natural. Entre a frase digitada pelo usuário e a saída de resposta deste módulo, ocorre a interação com o módulo de DLS

3.3.2 Módulo de Desambiguação Léxica de Sentido (DLS)

Este módulo interage com o módulo de Conversação quando a frase digitada pelo usuário contém uma palavra ambígua, e é de responsabilidade deste módulo fazer a desambiguação da palavra e indicar qual sentido é o pretendido para a palavra ambígua.

3.4 DELIMITAÇÕES

Este trabalho está delimitado a realizar:

- a desambiguação léxica de sentido de uma palavra ambígua definida previamente e que esteja dentro do domínio de conhecimento do chatterbot;
- guardar em memória o histórico da conversa com o usuário para extrair o contexto.

Neste trabalho não será realizado:

- tratamento da análise da conversação (abertura, desenvolvimento e fechamento);
- aprendizagem automática com o usuário;
- realização de conversas fora do escopo estabelecido como domínio de conhecimento do chatterbot;
- desambiguação de palavras não definidas;
- guardar informações a respeito do usuário, tais como nome e idade;
- não será utilizado nenhum léxico computacional como a WordNet.

4 DESENVOLVIMENTO

Neste capítulo, será demonstrado o desenvolvimento do chatterbot proposto para alcançar os objetivos deste trabalho. Aqui, serão relatados os problemas e dificuldades encontrados durante o percurso de seu desenvolvimento, assim como as soluções e as técnicas adotadas para o projeto.

Primeiramente, existe a definição da base de conhecimento e das palavras ambíguas, posteriormente, será descrito, então, o conjunto dos métodos utilizados para realizar a conversa com o usuário e, por fim, o conjunto de técnicas que farão parte do módulo de DLS.

4.1 DEFINIÇÃO DA BASE DE CONHECIMENTO

Foi estipulado que a base de conhecimento será sobre a cidade de Florianópolis, mais especificamente o turismo na cidade. Este assunto foi escolhido por se tratar de algo regional e de fácil entendimento. Também foi decido que a base será focada nas atrações turísticas culturais como museus e teatros, para, desse modo, evitar propagadas de restaurantes, hotéis e etc.

Segundo Coppin (2012, p. 211) "a base de conhecimento consiste em um conjunto de regras que representam o conhecimento que o sistema tem. A base de dados de fato representa entradas do sistema que são usadas para obter conclusões ou provocar ações"

4.1.1 Fatores para a escolha da base de conhecimento

Um fator importante de citar para a escolha do turismo em Florianópolis, como base de conhecimento para o chatterbot, é o fato de muitas das ruas, assim como museus e

teatros, ganharem seu nome devido a políticos e artistas famosos da região, como o Museu Cruz e Sousa, o museu Victor Meirelles, a Rua Pedro Ivo, a rua Hercílio-Luz e muitas outras que se podem descrever. Isso facilita a criação de frases que contenham ambiguidade para a criação da base de conhecimento, já que, ao entrar com o nome da pessoa, o usuário pode estar querendo se referir a pessoa em si ou ao local.

Além desse fator, existem os nomes de diversas praias que são substantivos comuns ou adjetivos, tais como “Solidão”, “Armação”, “Brava”, “Mole”, que fazem com que as frases se tornem ambíguas e auxiliem igualmente para a definição da base de conhecimento.

4.1.2 Definição das perguntas e respostas

As perguntas são definidas de forma manual. Primeiramente, é definida uma lista de perguntas que serão respondidas, e nessa lista são marcadas as palavras que necessitam de desambiguação e anotados os possíveis sentidos que essa palavra terá.

Para a pergunta ser recuperada na base de dados, será utilizada a canonização, processo explicado na seção 2.2.1.1 deste trabalho. Como exemplo, deste processo, pode ser citada como pergunta: “Onde fica a Hercílio-Luz?”, esta sendo a pergunta definida, a palavra ambígua seria Hercílio-Luz, e seus possíveis significados seriam “Ponte”, “Rua”, “Aeroporto” e “Pessoa”. A forma canônica desta frase ficaria “Onde + Hercílio-Luz”, formando assim duas palavras chaves: Onde e Hercílio-Luz. As palavras-chave são de extrema importância para a obtenção da resposta associada à pergunta entrada pelo usuário.

Abaixo, serão mostrados alguns exemplos de frases que serão tratadas pelo chatterbot.

Figura 11 – Quadro de exemplos de frases tratadas pelo sistema

Frase	**Palavra Ambígua**	**Possíveis Sentidos**	**Forma Canônica**
P: Onde fica a Hercílio Luz ?	Hercílio-Luz	Ponte, Rua, Aeroporto, Pessoa	Onde + Hercilio-Luz
P: Você sabe como fazer pra chegar em Ratones ?	Ratones	Bairro, Ilhas	Como + Chegar + Ratones
P: Quais os museus no centro da cidade ?	Sem ambiguidade	Sem ambiguidade	Museu + Centro + Cidade
P: Conte-me mais sobre a fortaleza de Anhatomirim	Sem ambiguidade	Sem ambiguidade	Mais + Fortaleza + Anhatomirim

Fonte: Elaboração do autor (2013).

Além das perguntas individuais, um caso interessante de ser relatado é quando uma pergunta depende da outra, como no seguinte exemplo:

Pergunta: Quais são as lagoas existentes ?

Resposta: As lagoas mais famosas são a lagoa do Peri e a lagoa da Conceição

Pergunta: Pode me falar mais sobre elas ?

Resposta: A lagoa do Peri se encontra no sul da ilha a 22km de distância do centro, e a lagoa da Conceição no leste, distante a 14km do centro da cidade.

Nesse exemplo, vê-se que a pergunta: “Pode me falar mais sobre elas?” está intimamente ligada a pergunta anterior “Quais são as lagoas existentes?”, dessa forma, utilizando da mesma ideia de definir uma palavra ambígua, pode-se definir a palavra “elas” como ambígua, que estaria se referindo às Lagoas, ou, em outros casos, às Trilhas, ou às Praias, dependendo do contexto.

Assim, como as perguntas, as respostas são definidas manualmente e, apesar da utilização da forma canônica diminuir a quantidade de perguntas para uma mesma resposta, é interessante pensar em formular a mesma pergunta, utilizando palavras que não estejam na primeira forma da pergunta, por isso muitas perguntas podem fazer referência a mesma resposta.

4.1.3 Categorias das perguntas

As perguntas deste trabalho foram separadas em categorias, entre elas pode-se listar:

- Cultura;
- Pessoas;

Dentro da categoria “Cultura” estão as perguntas referentes a teatros e museus.

Na categoria “Pessoas” existem as perguntas relacionadas às pessoas de renome na cidade, como artistas e políticos que fizeram parte da história da região.

4.2 MÓDULO DE CONVERSAÇÃO COM O USUÁRIO

Nesta seção é explicado o processo de obtenção das respostas para as entradas digitadas pelo usuário. Esse processo consiste em:

- Passar a pergunta do usuário pela lista de substituição e pela stoplist;
- Obtenção das palavras-chaves;
- Tratamento do tipo de pergunta;

4.2.1 Lista de Substituição

Esta é a primeira etapa do processo para encontrar a resposta desejada. Nesta parte, a pergunta do usuário passa pela verificação de cada palavra para substituir sinônimos ou palavras digitadas incorretas.

4.2.2 *Stoplist* e Obtenção das Palavras Chaves

Neste passo, são eliminadas as palavras desnecessárias (*stopwords*) através de uma lista contendo preposições, artigos e outras classificações de palavras.

Num segundo momento, são efetuadas verificações de palavras-chaves compostas, como, por exemplo, “Hercílio Luz” e “Santo Antônio de Lisboa”.

Ao fim desta verificação, obtém-se apenas as palavras que tenham importância para a pesquisa, ou seja, que são palavras-chaves, montando a forma canônica da entrada do usuário.

4.2.3 Classificação das perguntas quanto ao uso

Vê-se necessário dividir as perguntas quanto a sua necessidade de uso. Esta é uma classificação apenas para melhor entendimento de como funcionará a estrutura do sistema para a obtenção da resposta adequada.

As perguntas são divididas em:

- saudações;
- perguntas simples;
- perguntas ambíguas;
- perguntas simples com palavras ambíguas;
- perguntas sem respostas.

4.2.3.1 Saudação

As perguntas classificadas, nesta categoria, têm como objetivo serem respondidas como um convite para o usuário conversar com o chatterbot, dessa forma, chamando a atenção para iniciar uma conversa. Como exemplo de perguntas dessa categoria se tem:

"Olá!",

"Oi, como vai você?".

"Oi, tudo bem ?"

E, como possíveis respostas:

"Olá, estou bem, obrigado ! No que posso lhe ajudar sobre Floripa ?"

"Oi! Estou bem e você ?"

"Bem vindo! O que você gostaria de desfrutar em Florianópolis?"

4.2.3.2 Perguntas Simples

As perguntas simples não contêm ambiguidade e podem ser respondidas facilmente pelo chatterbot, caso sua resposta seja encontrada, sem que se faça necessário o acesso ao módulo de DLS.

Exemplos dessa pergunta são:

"Quantas praias existem em Florianópolis?"

"Por que a Ilha da Magia tem esse apelido ?"

"Quais os outros nomes que Floripa já teve ?"

Essas perguntas não necessitam acesso ao módulo de DLS, pois devido à inexistência de palavra ambígua, sua resposta pode ser encontrada diretamente, buscando pelas palavras-chaves.

4.2.3.3 Perguntas Ambíguas

São perguntas que contêm alguma palavra ambígua, fazendo com que seja necessário haver a desambiguação do sentido da palavra para obter uma resposta satisfatória. Essas são algumas das perguntas com as quais o sistema precisará acessar ao módulo de DLS para responder.

"Adorei a Brava, gostaria de uma parecida, qual você me recomenda ?"

"Adorei sua armação! Onde você conseguiu ?"

"Hoje, eu comi um dourado delicioso!"

Essas perguntas, muitas não serão necessariamente sobre Florianópolis, mas conterão algo em comum com a cidade ou com a cultura da região, como algum nome de praia, de bairro ou de algum peixe.

Nos exemplos citados, viu-se "Brava", que pode significar um adjetivo ou a praia de mesmo nome no norte da ilha. O segundo exemplo se utilizou como palavra ambígua "Armação" para falar de óculos, contudo, existe uma praia chamada "Armação" no Sul da Ilha, desse modo, sendo necessário fazer a desambiguação da palavra. No terceiro exemplo, há a palavra "Dourado", que pode representar um adjetivo ou o nome de um peixe.

4.2.3.4 Perguntas Simples com Palavras Ambíguas

Nessa categoria, enquadram-se as perguntas que contêm alguma palavra ambígua, mas que não deixam a frase com mais de um sentido. Apesar de não ser necessária a desambiguação da frase, é coerente com o objetivo deste projeto que o chatterbot informe que identificou a palavra ambígua e encontrou seu significado, mesmo que, para a pergunta em questão, exista uma ou nenhuma resposta, desde que a palavra ambígua esteja definida no domínio de conhecimento.

Como exemplo desta categoria, pode-se citar:

"Você já visitou a ponte Hercílio-Luz ?"

Como já é explícito na própria frase que se está falando da ponte, não há necessidade de desambiguação, porém, é importante mostrar ao usuário que o sistema identificou como “ponte” o sentido pretendido para a palavra-ambígua “Hercílio Luz”, para mostrar a capacidade do sistema.

4.2.3.5 Perguntas sem Respostas

Como o chatterbot tem um domínio de conhecimento limitado, perguntas feitas fora deste domínio ou aquelas para as quais de alguma forma o chatterbot não consiga encontrar uma resposta, são tratadas como desconhecidas e devem ser contornadas para que a conversa não fuja do assunto de conhecimento do chatterbot.

Como exemplo, pode-se citar qualquer pergunta que não esteja no escopo de domínio do chatterbot:

“Qual a velocidade da luz?”

“Quem descobriu o Brasil?”

Estas perguntas devem ser contornadas para que o chatterbot não pare de conversar com o usuário repentinamente.

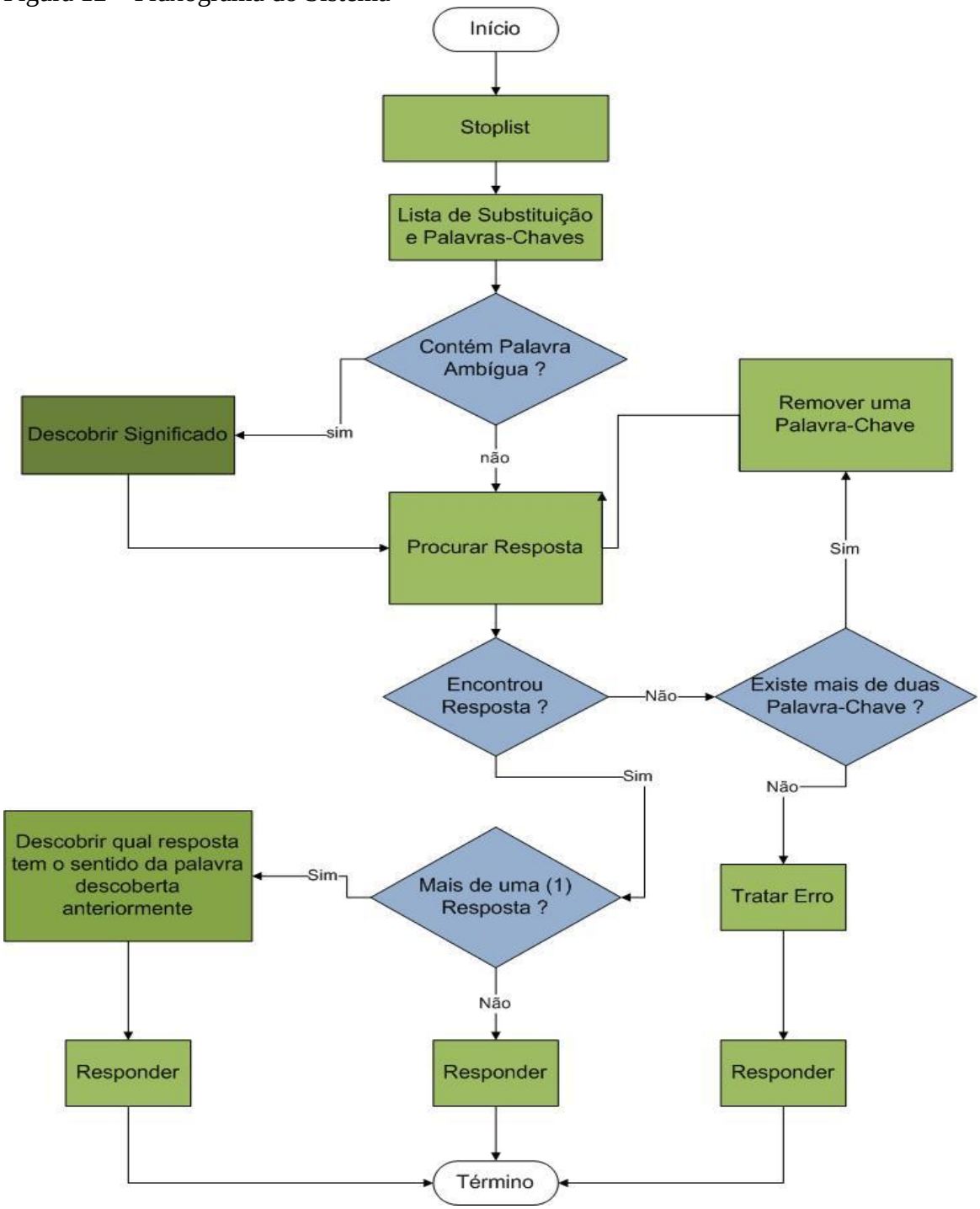
Como respostas de exemplo para essas perguntas, pode-se citar, respectivamente:

“Hmm.. boa pergunta.. não sei a resposta, posso lhe ajudar em algo mais ?”

“Não sei, mas posso lhe explicar como chegar a Lagoa da Conceição, pode ser?”

4.2.4 Fluxograma do funcionamento do chatterbot

Figura 12 – Fluxograma do Sistema



Fonte: Elaboração do autor (2013).

4.3 MÓDULO DE DESAMBIGUAÇÃO LÉXICA DE SENTIDO

Para realizar a desambiguação léxica de sentido da palavra ambígua da frase, foram escolhidas 4 (quatro) fontes de conhecimento:

- *stoplist*;
- *collocations*;
- *bag-of-words*;
- associação semântica de palavras.

A *Stoplist* é o primeiro passo do processo geral. Ela separa numa frase as partes significantes, excluindo o que é desnecessário. Sendo utilizada no início, é a única fonte de conhecimento que não está exclusivamente no módulo de DLS.

As outras fontes são utilizadas na desambiguação em si. Seus conceitos são semelhantes e sua montagem é realizada na definição das palavras ambíguas.

Ao se definir uma frase que contenha uma palavra ambígua, para essa palavra são definidos os possíveis sentidos para realizar sua desambiguação. Cada sentido agrupa um conjunto de palavras que ocorrem juntamente com a palavra ambígua no contexto de seu significado pretendido.

Todos os dados necessários para realizar a desambiguação de uma palavra, como a definição dos sentidos que ela pode ter, quais palavras co-ocorram próxima, ou que compartilham uma associação semântica, são montados previamente em conjunto com a base de conhecimento.

4.3.1 Associação Semântica de Palavras

Dentre essas palavras estão incluídas as palavras que compartilham uma mesma taxonomia (Associação Semântica de Palavras), como, por exemplo: Ao falar de comida e surgir a palavra ambígua “salmão” e, no mesmo contexto, estiverem as palavras: “tainha,

dourado, sardinha e atum", o módulo de DLS resultará que a palavra "salmão" está se referindo a "peixe", por compartilhar da taxonomia "Peixes" com essas outras palavras encontradas na conversa.

Para o desenvolvimento dessa fonte de conhecimento neste trabalho foram atribuídas palavras relacionadas a 'Pessoa' (governador, político, etc.) quando o sentido pretendido da palavra ambígua 'Hercílio Luz' fosse a pessoa, e palavras relacionadas a 'Pontes' (pontes, viadutos, construção, etc.) quando se referenciando à mesma palavra com o sentido do ponto turístico da cidade.

Abaixo segue um quadro exemplificando a utilização das palavras relacionadas.

Figura 13 – Quadro de exemplo de uma palavra ambígua

Palavra Ambígua	**Sentido**	**Palavras Relacionadas**
Hercílio Luz	Pessoa	Governador, político, mandato
	Ponte	Viaduto, construção, estrutura, monumento, obra, engenharia

Fonte: Elaboração do autor (2013).

4.3.2 Collocation

Além do compartilhamento de taxonomia, ainda são incluídas nesse conjunto de palavras aquelas que ocorram geralmente próxima às palavras com o sentido desejado (*Collocation*), como, por exemplo: Numa frase, ao encontrar a palavra ambígua: Solidão (Podendo se referir ao sentimento de estar só, ou a praia localizada no sul da ilha de Florianópolis") e contiver no texto da conversa as palavras: "Praia, Sul da Ilha, Surf, Trilha", o sistema resultará numa resposta para o sentido de "praia".

Ao definir a palavra 'Hercílio Luz' como ambígua, adiciona-se à lista de palavras relacionadas as palavras que co-ocorram com mais frequência junto com ela quando se referindo a ponte ou a pessoa.

Figura 14 – Quadro de Exemplo de uma palavra ambígua

Palavra Ambígua	**Sentido**	**Palavras Relacionadas (*Collocation*)**
Hercílio Luz	Pessoa	Quem, nascimento, morte, etc.
	Ponte	Ponto turístico, ponte velha, restauração, etc.

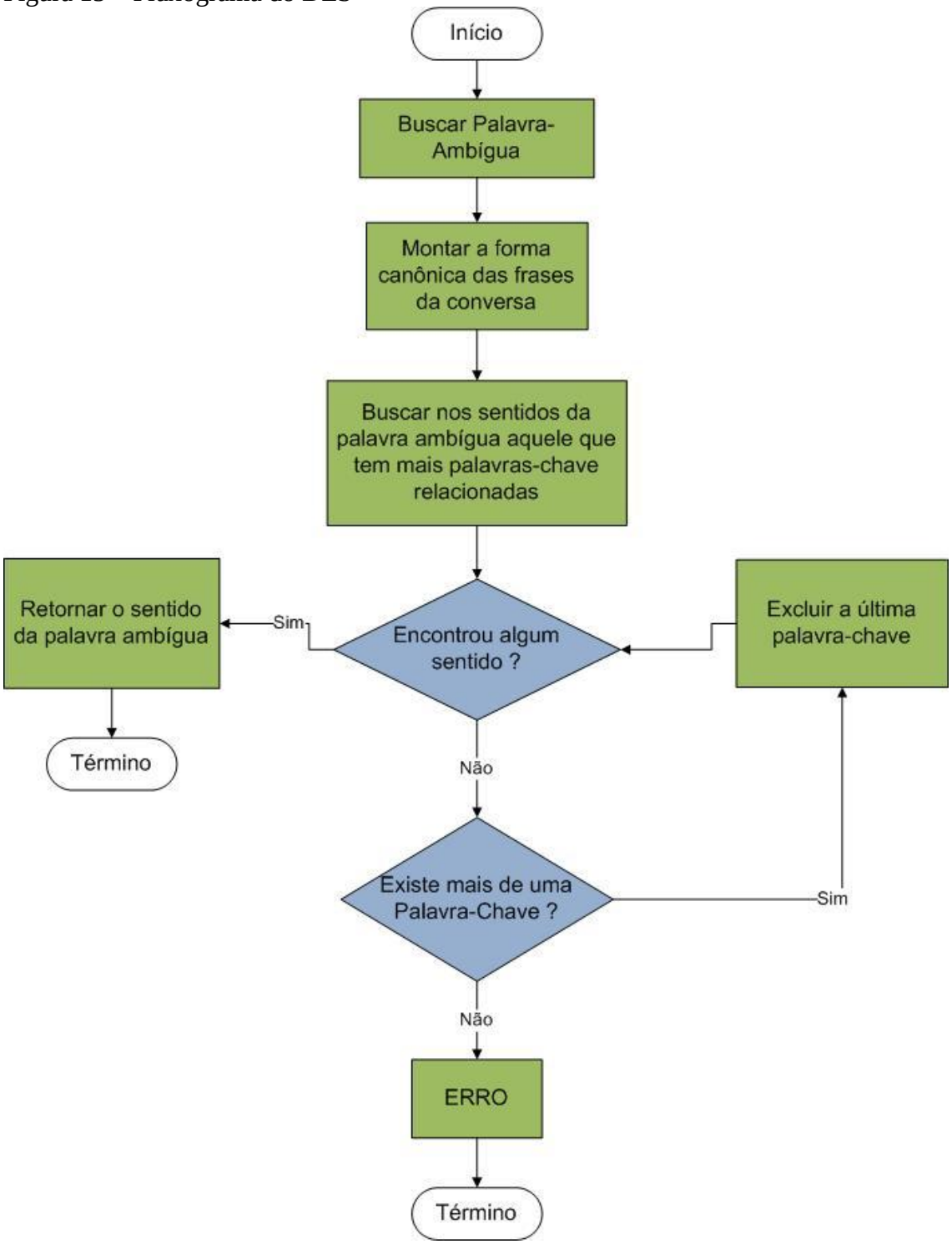
Fonte: Elaboração do autor (2013).

4.3.3 Bag-of-Words

Após reunir as palavras-chaves para procurar pelos sentidos da palavra ambígua, dá-se preferência às palavras que tenham ocorrido mais vezes no texto, utilizando, dessa maneira, a fonte *Bag-of-Words*. A lista de palavras-chaves a serem excluídas na pesquisa é a lista de *bag-of-words* ordenada crescentemente de acordo com a frequência de cada palavra na conversa.

4.3.4 Fluxograma do Módulo de DLS

Figura 15 – Fluxograma do DLS



Fonte: Elaboração do autor (2013).

4.4 REQUISITOS FUNCIONAIS

Como neste trabalho será desenvolvida uma versão de protótipo do sistema, vê-se necessário definir os requisitos funcionais que o sistema deve atender. Segundo Guedes (2011), o levantamento de requisitos aborda os requisitos funcionais e os não-funcionais e trabalha com o domínio do problema, determinando "o que" o software deve fazer e se é viável desenvolvê-lo.

Segundo Sommerville (2004) os requisitos funcionais "descrevem a funcionalidade ou os serviços que se espera que o sistema forneça". Abaixo, serão apresentados os requisitos funcionais do chatterbot desenvolvido neste projeto:

Figura 16 – Quadro dos Requisitos Funcionais

Requisito Funcional	**Descrição**
RF01	O sistema deve prover a funcionalidade de responder as perguntas digitadas pelo usuário em Linguagem Natural, desde que o assunto seja sobre Florianópolis.
RF02	O sistema deve prover a funcionalidade de identificar uma palavra ambígua numa frase digitada pelo usuário e realizar sua desambiguação.
RF03	O sistema deve prover a funcionalidade de iniciar uma conversa com usuário em linguagem natural com saudações.
RF04	O sistema deve prover a funcionalidade de terminar uma conversa com o usuário através de despedidas.

Fonte: Elaboração do autor (2013).

4.5 REQUISITOS NÃO-FUNCIONAIS

"Não dizem respeito diretamente às funções específicas fornecidas pelo sistema. Eles podem estar relacionados a propriedades de sistemas emergentes, como confiabilidade, tempo de resposta e espaço em disco" (SOMMERVILLE, 2004, p. 85).

A seguir, é mostrado um quadro com os requisitos não-funcionais do chatterbot.

Figura 17 – Quadro de Requisitos Não-Funcionais

Requisito Não Funcional	**Descrição**
RNF01	O computador do usuário deve estar conectado a web para acessar o sistema chatterbot.
RNF02	A interface gráfica do sistema deve ser semelhante a de programas de conversão (chat), para que o usuário se sinta de fato conversando com alguém.
RNF03	O sistema não responderá perguntas que estejam fora de sua base de conhecimento.

Fonte: Elaboração do autor (2013).

4.6 REGRAS DE NEGÓCIO

Segundo Guedes (2011) as regras de negócio são "políticas, normas e condições estabelecidas [...] que devem ser seguidas na execução de uma funcionalidade". A seguir são mostradas as regras de negócio do chatterbot.

Figura 18 – Quadro de Regras de Negócio do sistema

Requisito	**Descrição**
RN01	Desambiguação de Palavras: Para realizar a desambiguação léxica de sentido de uma palavra na frase digitada pelo usuário, deve-se recorrer a frases anteriores da conversa, extraindo as palavras-chave de cada frase para encontrar o contexto da conversa e, através do contexto, descobrir o sentido da palavra.
RN02	Resposta ao Usuário: Para o sistema responder ao usuário de forma mais coerente possível, o assunto da conversa deve ser sobre Florianópolis (Praias, trilhas, cultura). A frase entrada pelo usuário é tratada e são extraídas suas palavras-chaves, que serão pesquisadas na base de dados para encontrar uma resposta adequada. Caso exista na frase alguma palavra ambígua, é necessário, então, chamar o módulo de Desambiguação Léxica de Sentido, para realizar a desambiguação.
RN03	Saudações e Despedidas: A conversa com o usuário deve ser o mais natural possível, o que inclui realizar saudações de abertura e despedidas.

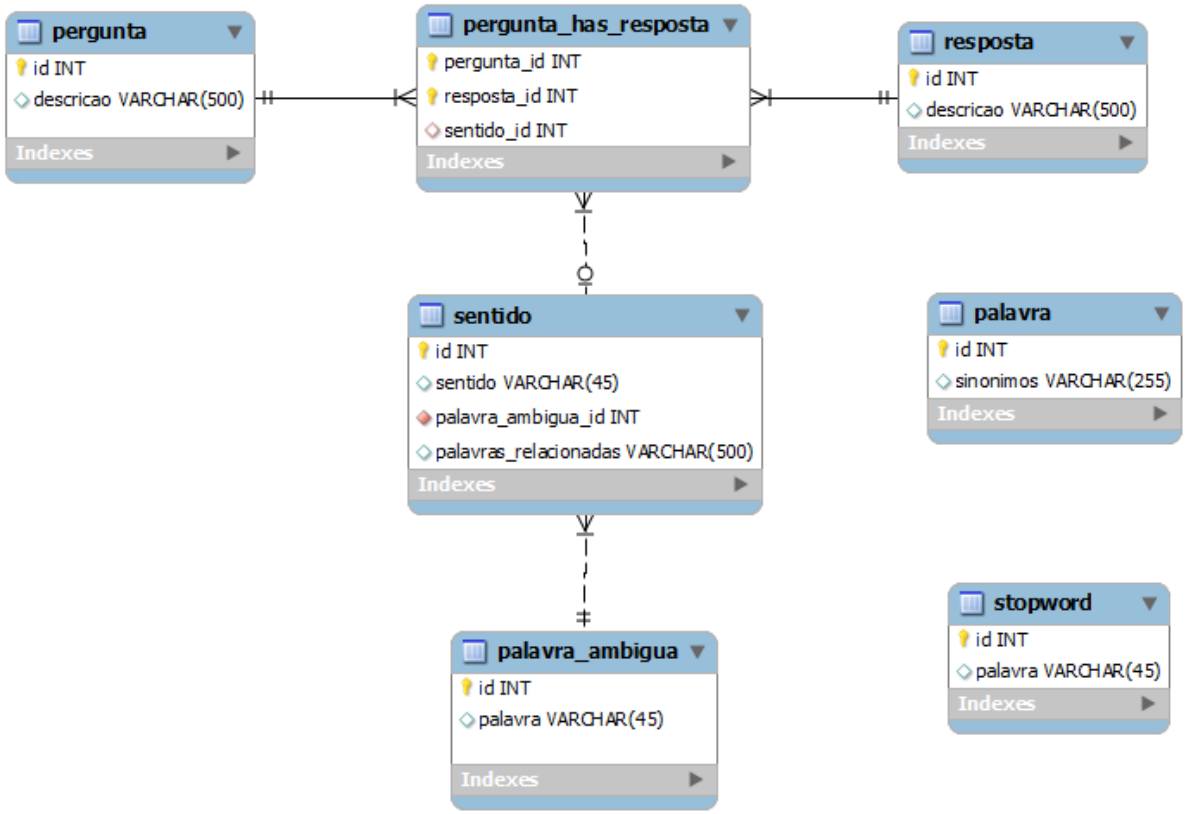
Fonte: Elaboração do autor (2013).

4.7 MODELO ENTIDADE-RELACIONAL

Silberschatz, Korth e Sudarshan (1999) afirmam quanto à modelagem de dados que é o processo para representar a visão que o usuário tem de seus dados e é uma das tarefas mais importantes no desenvolvimento de aplicações que utilizam banco de dados. O Modelo Entidade-Relacional, segundo os mesmos autores, "tem por base a percepção de o que mundo real é formado por um conjunto de objetos chamados entidades e pelo conjunto dos relacionamentos entre esses objetos".

Abaixo, é apresentado o Modelo Entidade-Relacional do chatterbot desenvolvido para este projeto.

Figura 19 – Modelagem Entidade-Relacional do chatterbot.



Fonte: Elaboração do autor (2013).

4.8 DIAGRAMA DE CLASSES

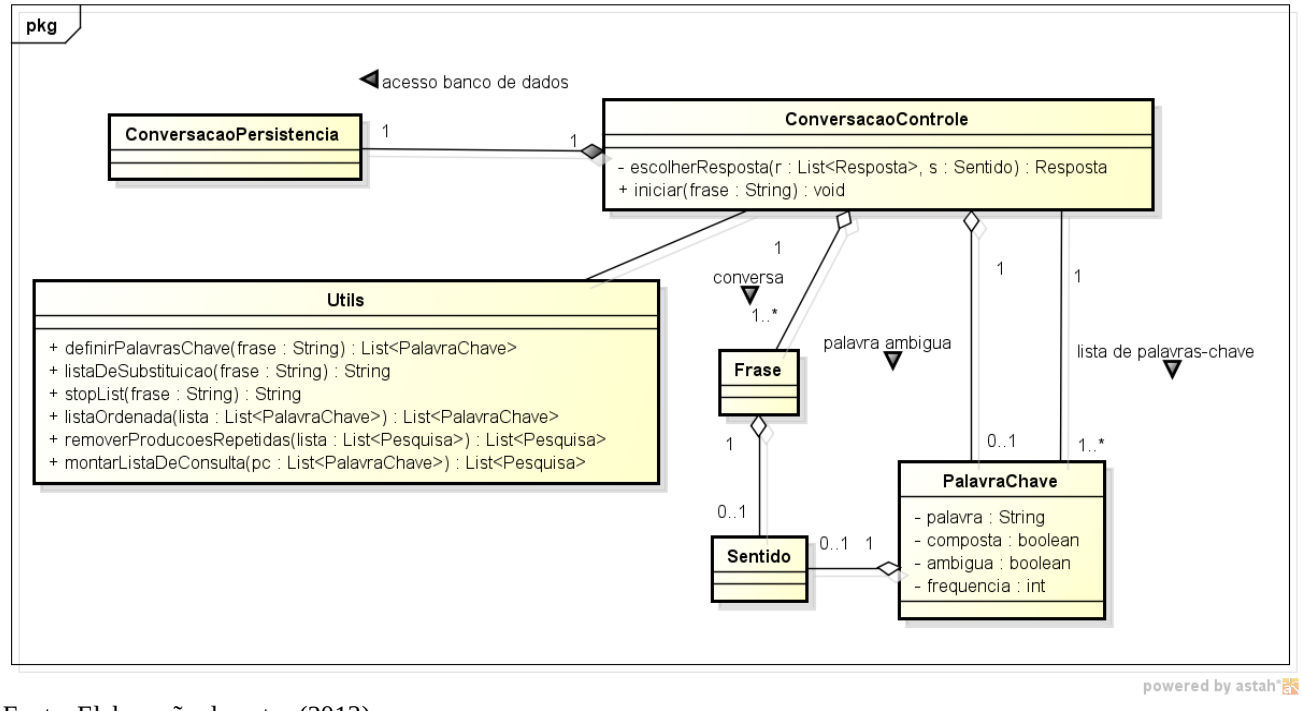
Para descrever o sistema de uma forma mais física escolheu-se utilizar o diagrama de classe da UML. O Diagrama de Classes tem como objetivo representar as classes e seus atributos assim como seus relacionamentos, apresentando uma visão estática de como as classes estão organizadas. (GUEDES, 2011)

Neste projeto serão apresentados dois diagramas de classes: O primeiro sendo do módulo de Conversação com o Usuário, e o segundo representando o módulo de Desambiguação do Sistema.

Para melhor visualização de cada diagrama, foi apresentado apenas o essencial.

4.8.1 Diagrama de Classes do Módulo de Conversação

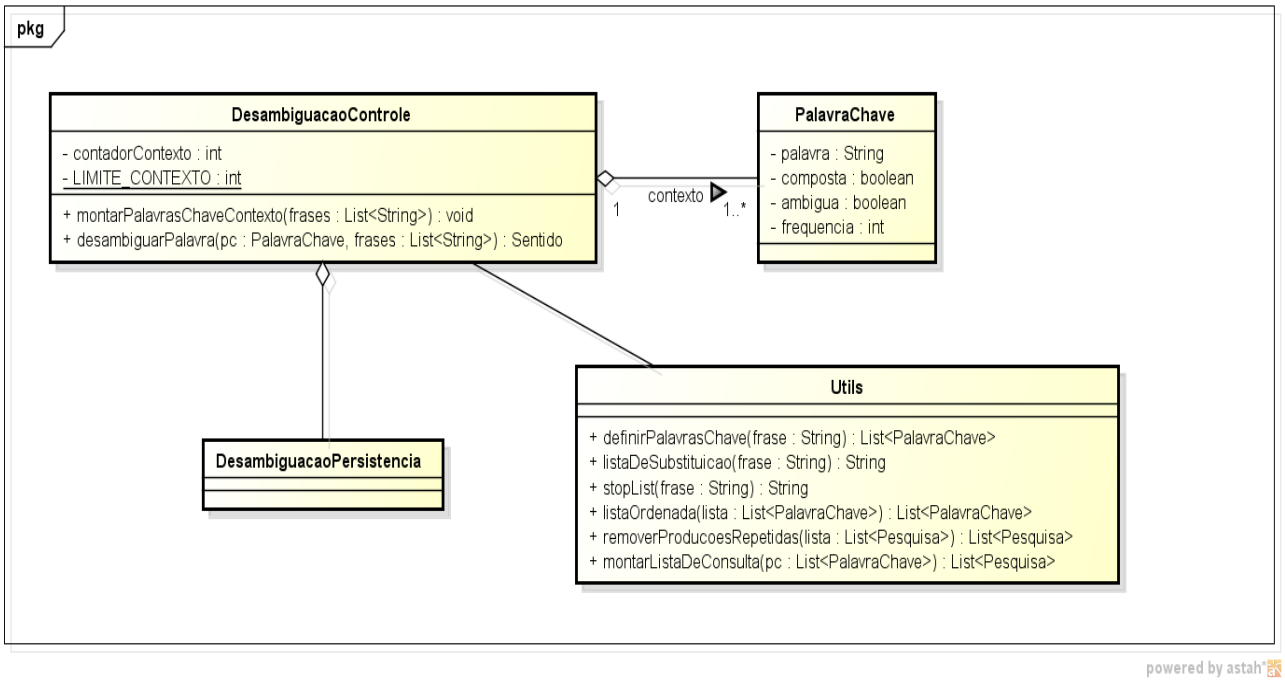
Figura 20 – Diagrama de Classes do Módulo de Conversação



Fonte: Elaboração do autor (2013).

4.8.2 Diagrama de Classes do Módulo de Desambiguação

Figura 21 – Diagrama de Classes do Módulo de Desambiguação



Fonte: Elaboração do autor (2013).

4.9 TECNOLOGIAS UTILIZADAS

Abaixo estão descritas as tecnologias utilizadas para o desenvolvimento do sistema chatterbot. O sistema foi construído utilizando a linguagem de programação Java, utilizando JSP (Java Server Pages) e Servlets, interagindo via web com um servidor web Tomcat, e colhendo informações de um banco de dados MySQL.

Após a descrição das tecnologias é apresentada uma imagem da interação entre as tecnologias.

4.9.1 Java

Java é uma linguagem de programação com o paradigma orientado a objetos desenvolvida pela *Sun Microsystems* na década de 1990. Foi apresentada oficialmente em 1995 e sua popularidade cresceu devido a sua utilização para criação de conteúdo dinâmico na web, apesar desse não ter sido o foco inicial para a linguagem (DEITEL; DEITEL, 2008).

A linguagem Java roda em cima de uma máquina virtual que faz parte da plataforma Java, possibilitando ao software desenvolvido não ficar atrelado a um sistema operacional, adquirindo dessa forma uma característica multi-plataforma (JAVA, 2013).

4.9.2 JSP e Servlets

Java Server Pages (JSP) e Servlets são tecnologias da família Java que possibilitam a construção de páginas web dinâmicas e independentes de plataforma. Os Servlets estendem a funcionalidade de um servidor web que servem páginas para um navegador utilizando o protocolo HTTP (DEITEL; DEITEL, 2008).

4.9.3 Apache Tomcat

O Apache Tomcat é um software *open-source*, que implementa as tecnologias JSP e Servlets. (APACHE Tomcat, 2013). Segundo Deitel e Deitel (2008) o Apache Tomcat inclui um servidor web para que possa ser utilizado como um container web.

4.9.4 MySQL

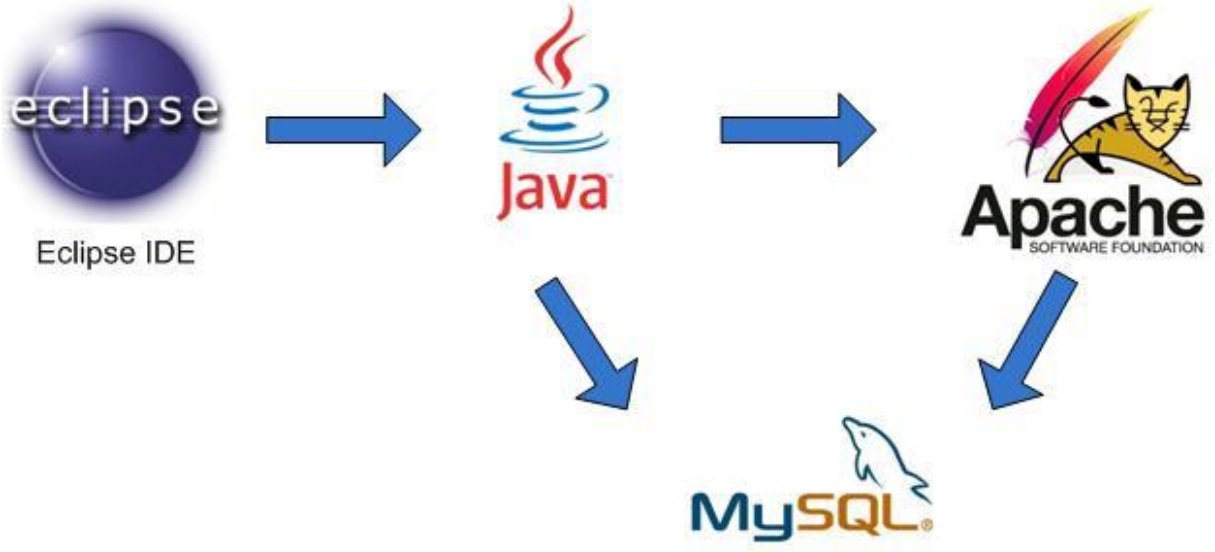
O MySQL é um gerenciador de banco de dados (SGBD) de código aberto, que utiliza da linguagem SQL (*Structured Query Language*) para realizar suas consultas. Distribuído pela Oracle (MySQL, 2013). Foi escolhido por ser compatível com diversas plataformas, ser de fácil utilização e ter bom desempenho nas consultas.

4.9.5 Eclipse IDE

Como ferramenta IDE (*Integrated Development Environment*) para o desenvolvimento do chatterbot foi utilizado o Eclipse. O Eclipse é um projeto da *Eclipse Fundation*, uma comunidade de indivíduos e organizações com o intuito de colaborar com a comunidade de software livre. Seus projetos são voltados para construir uma plataforma para auxiliar no desenvolvimento e gerenciamento de um software.

4.9.6 Interação entre as tecnologias utilizadas

Figura 22 – Interação entre as tecnologias do sistema



Fonte: Elaboração do autor (2013).

4.10 APRESENTAÇÃO DO SISTEMA

Abaixo são mostradas algumas imagens de conversas com o chatterbot. São apresentadas conversas onde as frases contenham a mesma palavra ambígua, mas com sentidos diferentes, e uma conversa onde não é necessário o tratamento da ambiguidade.

Figura 23 – Conversa com o chatterbot desambiguando "Victor Meirelles" como museu.

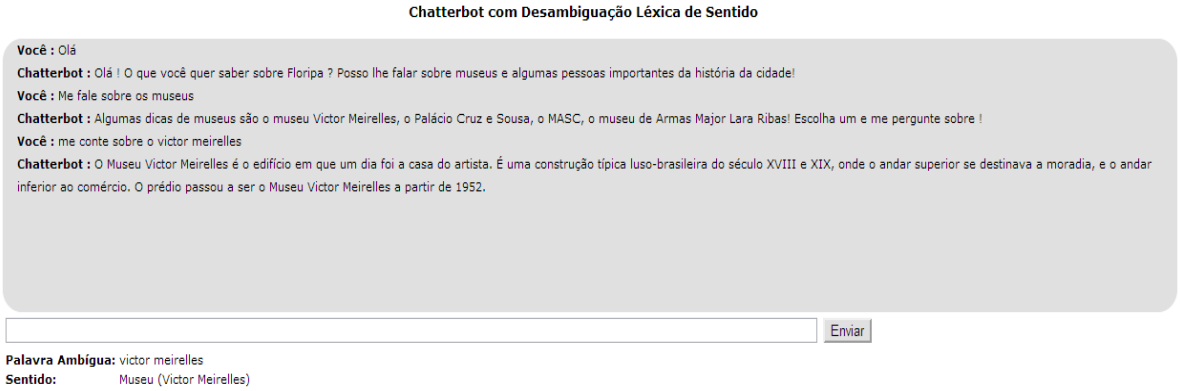
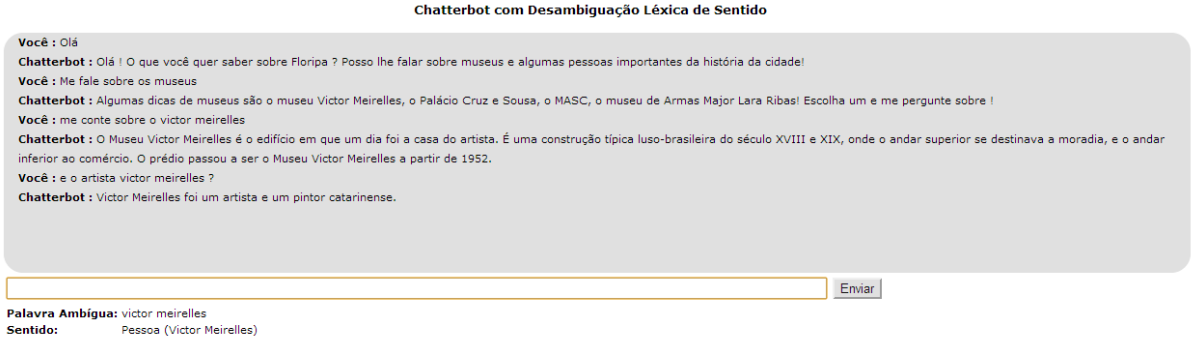
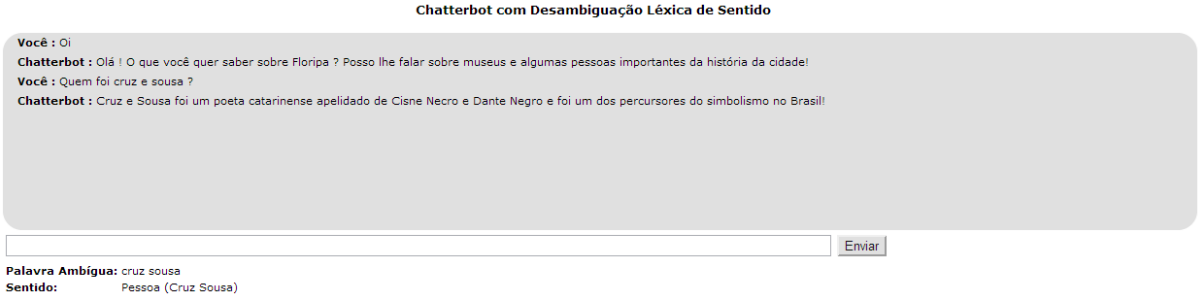


Figura 24 – Conversa com o chatterbot desambiguando “Victor Meirelles” como pessoa



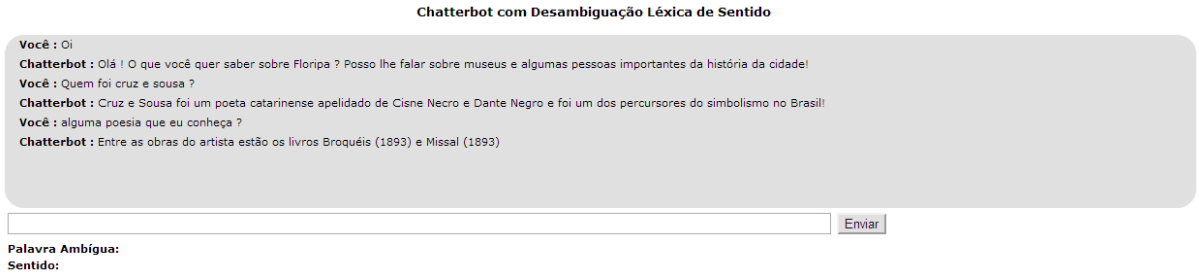
Fonte: Elaboração do autor (2013).

Figura 25 – Conversa com o chatterbot desambiguando “Cruz e Sousa” como pessoa



Fonte: Elaboração do autor (2013).

Figura 26 – Conversa com o chatterbot sem tratar ambiguidade



Fonte: Elaboração do autor (2013).

4.11 VALIDAÇÃO

Na validação do sistema pretende-se averiguar se o sistema atendeu aos objetivos definidos. Em uma conversa entre o usuário e o chaterbot, caso na entrada do usuário haja uma palavra ambígua definida previamente, o sistema deve fazer a identificação da palavra ambígua, e realizar sua desambiguação.

A validação foi realizada através de um formulário aplicado a 20 pessoas, utilizando da escala de Likert. Foi apresentado ao usuário o sistema e a lista de todas as palavras-ambiguas utilizadas, e solicitou-se que ele mantivesse uma conversa com o chatterbot sobre Florianópolis.

Para classificar as respostas dos usuários, utilizou-se a escala de Likert, codificando as respostas em: 1 = Nunca, 2 = Geralmente Não, 3 = As Vezes, 4 = Geralmente Sim, 5 = Sempre

O questionário verificou os seguintes aspectos em relação ao sistema chatterbot:

1) Facilidade de uso do sistema e ergonomia do sistema:
   - Você achou fácil a utilização e entendimento do sistema Chatterbot ?

2) Como o tempo de resposta é influenciado pela pesquisa das palavras, foi realizada a pergunta:
   - O tempo de resposta do chatterbot foi satisfatório?

3) Como o chatterbot é um sistema que simula uma conversação com o usuário, levou-se em consideração a coerência das respostas dadas pelo sistema. Para isso foi feita a seguinte pergunta:
   - A resposta obtida do chatterbot foi coerente com a pergunta realizada? Ou seja, a resposta se enquadra no assunto da conversa?

4) Para atingir o objetivo final é necessário que o sistema primeiro reconheça a palavra ambígua na frase do usuário, quanto a essa questão foram feitas as seguintes perguntas:
   - Entre as suas perguntas, alguma delas conteve uma das palavras-ambíguas apresentadas?
   - Em suas perguntas contendo uma palavra ambígua, ela foi identificada?

5) Após o reconhecimento da palavra ambígua, para atingir o objetivo é necessário buscar no contexto da conversa o assunto sendo tratado para realizar a desambiguação léxica de sentido. Neste ponto foram feitas as seguintes perguntas:

- Nas respostas obtidas pelo chatterbot através de uma pergunta contendo uma palavra ambígua, foi realizada a desambiguação da palavra? Ou seja, além de identificar a sua palavra como ambígua o sistema ainda apresentou o sentido pretendido?

- Caso o sistema não tenha encontrado a resposta para uma pergunta contendo uma palavra ambígua, mesmo assim a palavra foi desambiguada, ou seja, o sistema mostrou o sentido pretendido da palavra?

6) Quanto as pesquisas realizadas na área de Processamento de Linguagem Natural (PLN) e aos sistemas desenvolvidos com essa abordagem, foi realizada a seguinte pergunta:
- Você acha útil a interação homem-máquina através de linguagem natural?

4.11.1 Resultados da Validação

1) Quanto a facilidade de uso e ergonomia do sistema, para a pergunta:
Você achou fácil a utilização e entendimento do sistema Chatterbot ?
80% Respondeu Sempre;
20% Respondeu Geralmente Sim;

2) Quando ao tempo de resposta, utilizando a pergunta:
O tempo de resposta do chatterbot foi satisfatório?
30% Respondeu Sempre;
60% Respondeu Geralmente Sim;
10% Respondeu As vezes

3) Quanto a coerência na conversa, utilizando a pergunta:
A resposta obtida do chatterbot foi coerente com a pergunta realizada? Ou seja, a resposta se enquadra no assunto da conversa?

30% Respondeu Sempre;

50% Respondeu Geralmente Sim;

20% Respondeu As vezes;

4) Quanto ao reconhecimento da palavra ambígua, ao realizar as perguntas:

Entre as suas perguntas, alguma delas conteve uma das palavras-ambíguas apresentadas?

90% Sempre;

10% Geralmente Sim;

e:

Em suas perguntas contendo uma palavra ambígua, ela foi identificada?

80% Respondeu Sempre;

20% Respondeu Geralmente Sim;

5) Quando a desambiguação da palavra ambígua, as perguntas:

Nas respostas obtidas pelo chatterbot através de uma pergunta contendo uma palavra ambígua, foi realizada a desambiguação da palavra? Ou seja, além de identificar a sua palavra como ambígua o sistema ainda apresentou o sentido pretendido?

30% Respondeu Sempre;

30% Respondeu Geralmente Sim;

40% As vezes;

e:

Caso o sistema não tenha encontrado a resposta para uma pergunta contendo uma palavra ambígua, mesmo assim a palavra foi desambiguada, ou seja, o sistema mostrou o sentido pretendido da palavra?

80% Respondeu Sempre;

10% Respondeu Geralmente Sim;

10% Respondeu As vezes;

6) Quando ao estudo da área de Processamento de Lingugem Natural, a pergunta:

Você acha útil a interação homem-máquina através de linguagem natural?

40% Respondeu Sempre;

60% Respondeu Geralmente Sim;

5 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo apresenta as considerações finais da monografia e os trabalhos futuros sugeridos para dar continuidade a esta proposta e ao tema abordado.

5.1 CONCLUSÕES

A primeira fase desse projeto consistiu um levantamento da literatura existente na área. Com essa etapa do trabalho, concluiu-se que ao analisar a anatomia de um chatterbot, sua estrutura pode ser montada utilizando diversas técnicas como casamento de padrões, AIML e reformulação da frase digitada pelo usuário. O histórico dos chatterbots apresentou um paralelo claro com a evolução da computação e suas inovações, tais como as redes neurais e os avanços nos estudos em inteligência artificial.

Outro tema abordado na revisão bibliografia foi o estudo dos métodos de desambiguação léxica de sentido (DLS), com eles viu-se uma grande variedade de fontes de conhecimento para realizar esse tratamento.

Após a revisão bibliográfica deu-se início a etapa de desenvolvimento da proposta, para isso, construiu-se um protótipo de um chatterbot com DLS utilizando como fonte de conhecimento para tratamento da ambiguidade: *collocations,* associação semântica de palavras, *bag-of-words* e *stoplist*.

Para realizar a conversação com o usuário utilizou-se o casamento de padrão simples, extraindo palavras-chave da pergunta do usuário para buscar a resposta mais adequada na base de conhecimento.

A base de conhecimento se limitou a museus famosos da cidade e alguns pontos turísticos, assim como artistas e governantes que fizeram parte da história da cidade.

Este trabalho tem como diferencial a existência de um módulo de desambiguação léxica de sentido em um sistema chatterbot, resolvendo assim o problema ocorrido por palavras com mais de um sentido numa frase.

Dentre os problemas encontrados no desenvolvimento do protótipo, estão: as frases do usuário contendo apenas uma palavra, a demora no processamento de frases muito

extensas como entradas, e respostas equivocadas. Por não encontrar uma resposta específica para determinada pergunta, então o chatterbot diminui a quantidade de palavras-chave e encontra uma resposta, mas que não é a correta.

Perante o que foi apresentado no trabalho conclui-se que ainda há muitas brechas a serem trabalhadas e desenvolvidas para melhorar a interação homem-máquina através de sistemas de conversação.

5.2 TRABALHOS FUTUROS

Existe muito material a ser explorado nessa área de processamento de linguagem natural, tanto para o estudo da interação homem-máquina, quanto para o tratamento de textos como correção e tradução automática.

Como trabalhos futuros podem ser levantados o aprendizado automático de um chatterbot através da conversa com o usuário, utilizando algum método supervisionado ou não-supervisionado, redes neurais e raciocínio baseado em casos, podendo, cada vez mais, expandir os limites da base de conhecimento do sistema, e o tornando mais inteligente.

Também pode ser interessante desenvolver um analisador sintático da língua, para obter uma estrutura de frases na conversa, mais próxima da linguagem natural. No tratamento da desambiguação léxica, muitas outras fontes de conhecimento podem ser utilizadas para construir uma base de conhecimento e obter uma interação mais fiel com o usuário, tais como ontologia e redes semânticas, não apenas se limitando a aplicações chatterbots, mas podendo cobrir outros problemas, como a tradução automática de textos e a recuperação de informação textual.

Aspectos da língua como gírias, figuras de linguagem, reconhecimento de nomes próprios, e muitas peculiaridades de cada língua podem ser trabalhadas para melhorar os sistemas chatterbots. O tratamento da análise de conversação também se mostra muito útil, aproximando a conversa com o sistema daquelas do cotidiano, como em chats. Além disso, podem ser desenvolvidos trabalhos que façam o chatterbot não apenas conversar, mas também executar uma ação através de um pedido do usuário, como enviar um email, salvar uma anotação, e etc.

Com isso vê-se como é grande a gama de possibilidades a se estudar nessa área de processamento de linguagem natural.

REFERÊNCIAS

AGIRRE, E; MARTINEZ, D. **Knowledge Sources for Word Sense Disambiguation**. 2001, Proceedings of the Fourth International Conference TSD 2001, Plzen (Pilsen), Czech Republic, September 2001. Published in the Springer Verlag Lecture Notes in Computer Science series. Vaclav Matousek, Pavel Mautner, Roman Moucek, Karel Tauser (eds.), Disponível em < http://arxiv.org/ftp/cs/papers/0109/0109030.pdf > Acesso em: 05 set. 2012

APACHE Tomcat. Disponível em: <http://tomcat.apache.org/index.html> Acesso em: 03 maio 2013.

BARION, E. C. N; LAGO, D. Mineração de Textos. 2008, **Revista de Ciências Exatas e Tecnologia** Vol. III, Nº 3, Ano 2008.

CAFE, L. R; COMARELLA, M. A. **Chatterbot: Conceito, características, tipologia e construção**, *Informação & Sociedade*: Estudos, João Pessoa, v.18, n.2, p. 55-67, maio/ago. 2008 Disponível em: <http://www.brapci.ufpr.br/download.php?dd0=12346>
Acesso em: 15 ago. 2012

CANUTO, E. P. **VICTOR-P Um chatterbot com personalidade.** 2005, 68 f. Trabalho de Conclusão de Curso (Ciência da Computação). Universidade Federal de Pernambuco [Orientador: Flávia de Almeida Barros] Disponível em: <www.cin.ufpe.br/~tg/2005-1/epc.doc> Acesso em: 25 ago. 2012

CERVOO, A. L; BERVIAN, P.A. **Metodologia científica**. 4a edição. São Paulo : MAKRON Books, 1996

COPPIN, Ben. **Inteligência Artificial**. Rio de Janeiro : LTC, 2012

CORREA, Abel. **Robô de conversação aplicado a educação a distância como tutor inteligente.** Universidade Federal do Rio Grande do Sul. Monografia do curso de Especialização em Informática na Educação. Disponível em <http://chasqueweb.ufrgs.br/~abelcorrea/monografia_versao_final.pdf> Acesso em 08 ago. 2012

DEITEL, H. M; DEITEL; P. J. **Java Como Programar.** 6a Edição. São Paulo : Pearson Education do Brasil, 2008.

DIAS, A. D; HENN, G; SILVA, J. W. M. Tecnologia da Informação e Serviços de Referência Eletrônicos: Uma Proposta de Aplicação baseada em chatterbots e ontologias, 2007. Enc. Bibli: **R. Eletr. Bibliotecon**. Ci. Inf., Florianópolis, n.23, 1º sem. 2007. Disponível em: < http://www.periodicos.ufsc.br/index.php/eb/article/view/322/391> Acesso em: 31 ago. 2012

GALLIANO, A. G. **O método científico teoria e prática**. São Paulo : HARBRA ltda, 1986

GUEDES, G. T. A. **Uml 2 - Uma Abordagem Prática**, 2a Edição. São Paulo : Novatec Editora, 2009

HAHN, S et al. **Detecting Communication Acts in Email Messages**. 2008 Disponível em < http://davidscotthayden.com/pdf/hahn_uwssli.pdf> Acesso em: 29 set. 2012

HUTCHENS, J. L. **How To Pass the Turing Test by Cheating**, 1997. Disponível em < http://www.agent.ai/doc/upload/200403/hutc97_1.pdf> Acesso em: 08 ago. 2012

IDE, N; VERONIS; J. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. 1998, **Computational Linguistics**, Vol. 24, N. 1 40f. Disponível em <http://acl.ldc.upenn.edu/J/J98/J98-1001.pdf> Acesso em: 05 set. 2012

JACOB JUNIOR, A. F. L. **"Buti: um Companheiro Virtual baseado em Computação Afetiva para Auxiliar na Manutenção da Saúde Cardiovascular".** 2008. 103 f. Dissertação (mestrado) – Universidade Federal de Pernambuco, 2008. [Orientador: Flávia de Almeida Barros] Disponível em: <http://www.jacobjr.org/index.php?option=com_docman&task=doc_download&gid=4> Acesso em: 28 ago. 2012

HUTCHENS. J.L; ALDER, M. D. **Introducing MegaHal.** In D.M.W. Powers (ed.) NeMLaP3/CoNLL98 Workshop on Human Computer Conversation, ACL, pp 271-274, 1998

JAVA. Disponível em: <http://www.java.com/en/download/faq/whatis_java.xml> Acesso em 06 maio 2013

KRAUS, H. M; FERNANDES, A. **Desenvolvimento de um chatterbot para área Imobiliária integrando Raciocínio Baseado em Casos**. XII Simpósio de Informática e VII Mostra de Software da PUCRS Uruguaiana, 2007. Disponível em: < > Acesso em: 15 set. 2012

LAUREANO, E. A. G. C; **ConsultBot – Um Chatterbot Consultor para Ambientes Virtuais de Estudo na Internet** 1999. Trabalho de Graduação em Ciência da Computação – Universidade Federal de Pernambuco, Recife, Pernambuco, 1999. Disponível em: http://www.di.ufpe.br/~tg/1999-1/eagcl.doc> Acesso em: 06 ago. 2012

LAVEN, Simon; **What is a Chatterbot ?** Disponível em: <http://www.simonlaven.com/> Acesso em: 8 ago. 2012

LEITÃO, D.A. **Um Chatterbot para um ambiente de ensino de gerência de projetos.** 2004. Trabalho de Graduação (Bacharelado em Ciência da Computação) – Universidade Federal de Pernambuco, Recife, Pernambuco, 2004.

LEONHARDT, M. D. et al; **ELEKTRA: Um Chatterbot para Uso em Ambiente Educacional** Disponível em: <http://penta3.ufrgs.br/~elektra/info/artigos/chatterbot-Elektra%5B1%5D.PDF> Acesso em: 8 ago. 2012

LEFFA, V. J. **Textual Constraints In L2 Lexical Disambiguation**. System, England, v. 26, n. 2, p. 183-194, 1998.

LEONHARDT, M. D. **Doroty – Um chatterbot para Treinamento de Profissionais Atuantes no Gerenciamento de Redes de Computadores.** 2005, 110 f. Dissertação (mestrado) Universidade Federal do Rio Grande do Sul [Orientador: Liane Margarida Rockenbach Tarouco] Disponível em: <https://www.repositorioceme.ufrgs.br/bitstream/handle/10183/5659/000473673.pdf?sequence=1> Acesso em: 25 ago. 2012

LIMA, Rosana de Vilhena. **POLISSEMIA E/OU HOMONÍMIA**. Disponível em: <http://www.filologia.org.br/revista/36/10.htm>. Acesso em: 24 set. 2012.

MOURA, Silva de. **Piscochat: Chatterbot aplicado ao ensino de psiquiatria.** 2008. 126 f. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Centro Universitário Feevale, 2008. [Orientador: Prof. Msc. Alexandre Oliveira Zamberlam] Disponível em: <http://tconline.feevale.br/tc/files/0002_1531.pdf> Acesso em: 8 ago. 2012

MAULDIN, M. **Chatterbots, Tinymuds, And The Turing Text: Entering The Loebner Prize Competition.** 1994. Disponível em: <http://robot-club.com/lti/pub/aaai94.html>. Acesso em 28 ago. 2012

MARCUSCHI, Luiz Antonio. **Análise da Conversação**. 5. ed. São Paulo: Editora Ática, 2003. (Princípios).

METZLER JR, D. A. **Beyond bags of words: Effectively Modeling Dependence and features in information retrieval**. 2007, Disponível em: < http://maroo.cs.umass.edu/pub/web/getpdf.php?id=779 > Acesso em: 29 set. 2012

MySQL. Disponível em <http://www.mysql.com/about/> Acesso em: 03 maio 2013

NAVIGLI, R.Word sense disambiguation: A survey. 2009,
ACM Computing Surveys, Vol. 41, No. 2, Article 10, 69f. Disponível em: < http://promethee.philo.ulg.ac.be/engdep1/download/bacIII/ACM_Survey_2009_Navigli.pdf > Acesso em 05 set. 2012

NEVES, A M. M.; BARROS, F. A. **iAIML: Um Mecanismo para Tratamento de Intenção em Chatterbots,** In: ENIA, 18., 2005, São Leopoldo. Anais... São Leopoldo, 2005. p.1032-1041. Disponível em < http://www.unisinos.br/_diversos/congresso/sbc2005/_dados/anais/pdf/arq0150.pdf> Acesso em: 31 ago. 2012

OLIVEIRA, G. S; LOPES, R. S; TRIDA, G. C; SANTANA; L. B. **UAMBot – Uma proposta para chatterbot especialista**, 2009. Trabalho de Graduação (Graduação em Sistemas de Informação) - Universidade Anhembi Morumbi, São Paulo, São Paulo, 2009.

OLIVEIRA NETO; S. F. de. **Abordagem Automática para Criação de Córpus Etiquetados com Sentidos para Desambigüação Lexical de Sentido na Tradução Inglês –**

Português. 2004. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação com ênfase em Análise de Sistema) - Centro Universitário de Araraquara – UNIARA [Orientador: Prof. Lucia Specia]

PRIBERAM, **Dicionário Priberam da Língua Portuguesa.** Disponível em < http://www.priberam.pt/dlpo/default.aspx?pal=polissemia> Acesso em 07 out. 2012

PRIMO, Alex; COELHO, Luciano Roth. **Comunicação e inteligência artificial: interagindo com a robô de conversação Cybelle**. In: MOTTA, L. G. M. et al. (Eds.). Estratégias e culturas da comunicação ed.Brasília. Brasília: Editora Universidade de Brasília, 2002. p. 83-106.

PRIMO, Alex Fernando Teixeira; COELHO, Luciano Roth; PAIM, Marcos Flávio Rodrigues et al. **O uso de chatterbots na educação à distância**. Porto Alegre: 2000. Universidade Federal do Rio Grande do Sul, UFRGS. Disponível em: <http://www.nied.unicamp.br/oea/mat/chatterbots_lec.pdf>. Acesso em 20 ago. 2012

PREISS, J**. Probabilistic word sense disambiguation Analysis and techniques for combining knowledge sources**. 2006, Technical Report n. 673. Disponível em < > Acesso em 06 out. 2012

ROTHERMEL. A; DOMINGUES, M. J. C. S**. Maria: Um chatterbot desenvolvido para os estudantes da disciplina "Métodos e Técnicas de Pesquisa em Administração"**. Trabalho apresentado ao IV Simpósio de Excelência em Gestão e Tecnologia, Resende, 2007. Disponível em: <http://www.aedb.br/anais-seget07/arquivos/ti/923_chatterbot.pdf> Acesso em: 27 ago. 2012

RUSSEL, S. J; NORVIG, P. **Inteligência Artificial**. Rio de Janeiro : Elsevier, 2004.

RICH, E; KNIGHT, K; **Inteligência Artificial**. São Paulo : Makron Books, 2993

SAYÃO, M; **Verificação e validação em requisitos: processamento da linguagem natural e agentes.** 2007, 205f. Teste (Doutorado em Informática) – Pontificia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. [Orientador: prof. Julio Cesar Sampaio do Prado Leite] Disponível : < > Acesso em: 29 set. 2012

SALLES, J. F; JOU, G. I; STEIN, L. M. O paradigma de priming semântico na investigação do processamento de leitura de palavras, 2007. **Interação em Psicologia**, Curitiba, jan./jun. 2007, (11) 1, p. 71-80. Disponível em < > Acesso em 06 out. 2012

SARAMENTO, L. **Agrupamento de contextos de palavras polisémicas.** 2006, Relatório Técnico ProDEI - FEUP (não publicado). 2006 Disponível em: <http://paginas.fe.up.pt/~las/conteudo/pub/pln/prodei/ec_luis_sarmento.pdf> Acesso em: 24 set. 2012

SANTOS, A. R. **Metodologia Científica a construção do conhecimento**. 5ª edição. Rio de Janeiro : DP&A, 2002

SILBERSCHATZ, A; KORTH, F. H; SUDARSHAN, S. **Sistema de Banco de Dados**, 3a Edição. São Paulo : MAKRON Books, 1999

SOMMERVILLE, Ian. **Engenharia de Software**. São Paulo : Addison Wesley,2004

SPECIA, L; NUNES, M. G. V. **Desambiguação Lexical Automática de Sentido: Um panorama.** 2004, Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional – NILC – ICMC - USP. Disponível em <http://clg.wlv.ac.uk/papers/Specia_NILC-TR-04-08.pdf> Acesso em: 24 set. 2012

TURING, A. M. **Computing machinery and intelligence**. Mind, v.59, n.236, p.433-460. 1950. Disponível em: <http://www.loebner.net/Prizef/TuringArticle.html>. Acesso em: 08 ago. 2012.

VIEIRA, Renata ; LIMA, Vera Lúcia Strube de . JAIA/Lingüística computacional: princípios e aplicações. In: Ana Teresa Martins; Díbio Leandro Borges. (Org.). *As Tecnologias da informação e a questão social:* anais. .1 ed. Fortaleza: SBC, 2001, v. 3, p. 47-88. Disponível em: <http://www.inf.unioeste.br/~jorge/MESTRADOS/LETRAS%20-%20MECANISMOS%20DO%20FUNCIONAMENTO%20DA%20LINGUAGEM%20-%20PROCESSAMENTO%20DA%20LINGUAGEM%20NATURAL/ARTIGOS%20INTERESSANTES/lingu%edstica%20computacional.pdf> Acesso em: 27 set. 2012

WALLACE, R. **AIML Overview**, 2012. Disponível em: <http://www.pandorabots.com/pandora/pics/wallaceaimltutorial.html> Acesso em: 28 ago. 2012

WAZLAWICK, R. S; CASTANHO, C. L. O. **A Avaliação do Uso de Chatterbots no Ensino Através de uma Ferramenta de Autoria**, 2002, XIII Simpósio Brasileiro de Informática na Educação – SBIE – UNISINOS 2002. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/160/146> Acesso Em: 30 ago. 2012

What is the Loebner Prize? Disponível em: <http://www.loebner.net/Prizef/loebner-prize.html>

WILKS, Y; CATIZONE, R; TURUNEN, M. **Dialogue Management** *Companions Consortium:* State of The Art papers 2, jan. 2006 Disponível em < http://www.peachbit.org/sites/peachbit.org/files/SoA_papers.pdf> Acesso em 10 ago. 2012

WEIZENBAUM, J. **ELIZA – A Computer Program For The Study of Natural Language Communication Between Man and Machine**, 1966. Communications of the ACM, v. 9 n. 1 Janeiro 1966: 36-35 Disponível em: < http://ai.vancouver.wsu.edu/~nwdcsd/wiki/images/4/47/ElizaScript.pdf> Acesso em 10 set. 2012