Chatterbots em ambientes de aprendizagem – uma proposta para a construção de bases de conhecimento

Sérgio Teixeira[1,2], Thiago Bortolo Ramiro[2], Elias de Oliveira[2], Crediné Silva de Menezes[2]

[1]Faculdade Salesiana de Vitória (UNISALES)
Av. Vitória, 950 – 29.040-330 – Vitória – ES – Brasil

[2]Mestrado em Informática – Universidade Federal do Espírito Santo (UFES)
Av. Fernando Ferrari, s/n, 29.060-900 – Vitória – ES – Brasil

sergio@multicast.com.br, thiagobortolo@yahoo.com.br,
elias@npd.ufes.br, credine@inf.ufes.br

Abstract. *This paper presents a proposal for building chatterbots knowledge base using the ALICE technology. This knowledge base is created based upon the dialog patterns existent on a Linguistics Corpus. The patterns are classified by the Latent Semantic Indexing method and rearranged by a specialist when necessary. As a case of study it is proposed the use of the Tuxbot, a simple chatterbot that provides the user of Linux with answers related to their doubts.*

Resumo. *Este artigo apresenta uma proposta de construção de bases de conhecimento de um chatterbot utilizando a tecnologia ALICE. A criação das bases de conhecimento é feita através da identificação de padrões de conversação baseados na Lingüística de Corpus. Os padrões são categorizados através do método Latent Semantic Indexing e manipulados por especialistas. Como estudo de caso, é proposto o Tuxbot, chatterbot que responde perguntas sobre dúvidas de usuários do Linux.*

Keywords. *Chatterbot, Lingüística de corpus and Latent Semantic Indexing.*

1. Introdução

A ausência das interações síncronas, presenciais ou não, é o principal fator de desestímulo dos alunos de cursos virtuais. Assim, o grande desafio para os pesquisadores da área é o desenvolvimento de soluções que busquem minimizar o sentimento de isolamento. Apesar da existência de estratégias e ferramentas com o objetivo de minimizar as dificuldades, ainda existem muitas barreiras geradas pela falta de habilidade dos alunos na utilização das tecnologias disponibilizadas nos ambientes virtuais de aprendizagem.

O uso de Chatterbots está sendo experimentado como solução alternativa aos *frequently ask questions* (FAQs). Por ser uma alternativa recente, requer estudos mais profundos. Esse trabalho propõe uma estratégia de criação de uma base de conhecimento de um Chatterbot que esclarece dúvidas de aprendizes em linguagem natural sobre um determinado domínio do conhecimento. A estratégia adotada está apoiada na modelagem do conhecimento através da recuperação de padrões[1] de conversação baseada na lingüística de corpus[2][3], na categorização desses padrões com a utilização do método *Latent Semantic Indexing*[4] (LSI) e na conversão supervisionada desses padrões em conhecimento do Chatterbot. Como estudo de caso foi desenvolvido um Chatterbot denominado Tuxbot que tem o objetivo de esclarecer dúvidas sobre o Linux.

A seção seguinte apresenta o que é e como se faz um Chatterbot e a arquitetura de um Chatterbot baseado na tecnologia ALICE. A seção três apresenta uma metodologia para a construção das bases de conhecimento. A seção quatro apresenta o Tuxbot – um estudo de caso na utilização de Chatterbot em ambientes de aprendizagem e a utilização da metodologia proposta. A seção cinco apresenta as considerações finais e a seção seis apresenta as referências bibliográficas.

2. Chatterbot. O que é e como se faz?

Chatterbot é um programa de computador que tenta simular um ser humano na conversação com as pessoas. O objetivo é responder às perguntas de tal forma que as pessoas tenham a impressão de estar conversando com outra pessoa e não com um programa de computador. Após o envio de perguntas em linguagem natural, o programa consulta uma base de conhecimento e em seguida fornece uma resposta que tenta imitar o comportamento humano [Teixeira 2003].

---

[1] Padrão: conjunto de palavras que poderão auxiliar na identificação de formas de elaboração de perguntas ou respostas.

[2] Lingüística de *corpus:* Área que estuda a coleta e exploração de conjuntos de dados lingüísticos textuais coletados criteriosamente.

[3] *Corpus*: conjunto de textos autênticos, em linguagem natural, escritos por falantes nativos, com o propósito específico de pesquisa lingüística. Os textos devem ser selecionados de forma aleatória, obedecendo a um conjunto de regras pré-estabelecidas.

[4] Baseado no modelo vetorial, esse método faz o mapeamento dos vetores de documentos e consultas em um espaço vetorial reduzido associado aos conceitos. Esse modelo tende a realçar as relações semânticas ocultas entre termos e documentos. [Deerwester 1990].

Dentre as tecnologias de Chatterbot existentes, merece destaque a tecnologia ALICE, a qual é baseada em Extensible Markup Language (XML). ALICE é uma tecnologia de desenvolvimento que se baseia na interpretação de bases de conhecimento escritas na linguagem Artificial Intelligence Markup Language (AIML) [Alice 2005].

As bases em AIML guardam os <*patterns*[5]> em uma estrutura de árvore gerenciada por um objeto chamado Graphmaster. Graphmaster fica residente em memória, possibilitando uma rápida e eficiente busca de <*patterns*>. O quadro 1 apresenta um exemplo de unidade de conhecimento.

```
<category>
<pattern>Como configurar um placa de som ISA no Linux?</pattern>
<template>
O padrão é a utilização da IRQ 5, DMA1, DMA16, I/O 0x220, 0x330, 0x388 para PCM e MIDI. A configuração de uma placa Plug-and-Play é descrita na seção 3.4.4 do Foca GNU/Linux e de uma placa com Jumpers, na seção 3.4.1 do Foca GNU/Linux.
</template>
</category>
```

Quadro 1. Unidade de conhecimento escrita na linguagem AIML

Experiências com a ALICE indicam que aproximadamente 2.000 palavras atendem 95% das opções escolhidas pelas pessoas como a primeira palavra no início de uma frase, a partir da segunda as opções diminuem bastante. Com aproximadamente 41.000 unidades de conhecimento é possível estabelecer um bom diálogo.

2.1 Arquitetura de um Chatterbot baseado na Tecnologia ALICE

A figura 1 apresenta uma visão geral do funcionamento do ALICE. A primeira etapa representa o encaminhamento da pergunta. Em seguida, o sistema realiza uma série de passos até que a pergunta fique pronta para a busca na base de conhecimento. Na segunda etapa, será feita a busca da pergunta nas bases de conhecimento contida nos arquivos AIML. Após a localização o sistema apresenta a resposta cadastrada.

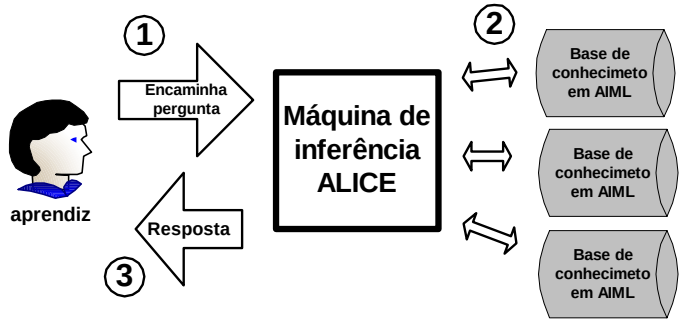

Figura 1. Arquitetura de um Chatterbot baseado na tecnologia ALICE.

Para que o Chatterbot seja capaz de responder as perguntas dos padrões cadastrados nas bases de conhecimento em AIML, é preciso executar uma rotina que lê os arquivos AIML e, em seguida, carrega-os na memória, de acordo com a ordem em que eles são

[5] Equivale a pergunta ou parte dela que fica armazenada na base de conhecimento.

lidos. Conforme a recomendação descrita na documentação da ALICE, as bases de conhecimento dever ser carregadas em ordem alfabética, sendo que, as bases de conhecimento que possuem perguntas que iniciam com o coringa "_", serão carregadas primeiro. Esse procedimento é importante para evitar problemas na busca de respostas e para auxiliar o *botmaster*[6] na manutenção das bases de conhecimento.

O funcionamento do mecanismo de inferência se dá através dos seguintes passos:

1º. PASSO: Simplificação e normalização

Após o encaminhamento da pergunta o sistema fará um tratamento no texto digitado para que a sentença fale a "mesma língua" da base de conhecimento. Através da identificação de palavras-chaves o texto será convertido para o formato mais apropriado para a busca na base de conhecimento. Por exemplo, quando o aluno digitar "Foca Gnu/Linux" o sistema transformará essas palavras em "Guia Foca", ou quando for digitada a palavra "usar", "utilizar" ou "uso" o sistema converterá para "como funciona" e caso seja digitado "o comando" o sistema converterá para "comando". Exemplo de simplificação e normalização de uma pergunta feita ao chatterbot:

Aluno: Favor informar como utilizo o comando find?
A pergunta será substituída por: Favor informar como como funciona comando find.

2º. PASSO: Padronização

É necessário padronizar a forma como a sentença será encaminhada ao passo seguinte. Serão identificadas as palavras-chaves que devem ser encaminhadas. O que não interessa será desprezando. O asterisco representa uma ou mais palavras.

Toda pergunta do tipo: "* como funciona * x" (leia-se: uma ou mais palavras, seguida da expressão "como funciona", seguida de uma ou mais palavras e encerrada por uma constante X) Será transformada em: "como funciona x"

Utilizando o exemplo do passo um, o sistema irá padronizar a sentença: "Favor informar como como funciona comando find" em "como funciona find"

3º. PASSO: Busca na base de conhecimento

Com a pergunta devidamente padronizada, o sistema fará a busca na base de conhecimento. Caso seja localizada, o chatterbot fornecerá a resposta.

3. Construção das bases de conhecimento

Um dos principais problemas encontrados na maioria dos Chatterbots é a falta de uma base de conhecimento consistente que permita ao Chatterbot responder perguntas diferentes daquelas consideradas óbvias sobre um determinado assunto.

O grande desafio dos pesquisadores da área é identificar a forma como as pessoas elaboram perguntas sobre dúvidas de um determinado domínio do conhecimento na Internet. Algumas técnicas baseadas na psicologia rogeriana[7] conseguem estabelecer um

[6] Pessoa que administra a base de conhecimento do chatterbot

[7] A psicologia rogeriana utiliza uma abordagem centrada na pessoa. Ela parte do pressuposto de que as respostas para as pergunta do paciente estão nele mesmo

diálogo, entretanto, elas não são capazes de responder perguntas específicas que podem ser elaboradas de diversas formas [Weizenbaum 1966].

Na busca por alternativas que agilizem a construção dessas bases de conhecimento formulou-se uma proposta constituída a partir das seguintes etapas:

1. Definir os padrões a serem recuperados de acordo com uma metodologia baseada na lingüística de *corpus*. Através da observação dos resultados gerados com diversos tamanhos de padrões, verificou-se que um conjunto de 16 palavras forma uma boa estrutura para auxiliar na identificação do padrão de conversação adequado a especificidade desse trabalho;
2. Definir os objetivos e utilidade dos *corpora*[8];
3. Definir a categorização e o tratamento que será executado nos arquivos;
4. Definir o processo de manipulação que será feito pelos especialistas no domínio do conhecimento e qual o layout dos arquivos após esse processo;

Nas subseções a seguir apresentamos os diversos elementos da proposta.

3.1 Como definir um padrão de conversação

A lingüística de corpus ocupa-se da coleta e da exploração de corpora, ou conjunto de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. Um marco na Lingüística de *Corpus* foi o lançamento, em 1964, do *corpus brown*, primeiro corpus lingüístico criado através da utilização de cartões perfurados. [Sardinha 2004].

Um padrão é definido como um conjunto de palavras que ocorrem em uma determinada freqüência, tendo em comum uma palavra específica e um significado associado.

A identificação de padrões busca a resposta para os seguintes questionamentos:

- Quais os padrões lexicais dos quais a palavra faz parte?
- A palavra se associa regularmente com outros sentidos específicos?
- Em quais estruturas ela aparece?
- Há uma correlação entre o uso/sentido da palavra e as estruturas das quais ela participa?
- A palavra está associada a uma certa posição na organização textual?

O fenômeno da colocação é o mais tradicionalmente enfocado no estudo de corpus. De acordo com Firth [Firth 1957], uma palavra deve ser julgada por sua companhia. Existem três definições principais, segundo Partington [Partington 1998]:

1. Textual: Colocação é a ocorrência de duas ou mais palavras distantes um pequeno espaço de texto uma das outras;

[8] Plural de *corpus* em latim

2. Psicológica: O sentido colocacional consiste nas associações que uma palavra faz por conta dos sentidos das outras palavras que tendem a ocorrer no seu ambiente.
3. Estatística: A relação que um item lexical tem com itens que aparecem com probabilidade significativa no seu contexto.

3.2 Objetivos e utilidade dos corpora propostos

Para criar as bases de conhecimento em AIML são necessários os seguintes corpora:

1. **Corpora de perguntas genéricas**: composto por textos que contenham diálogo, principalmente o diálogo característico de ambiente virtual, escritos por falantes nativos da língua portuguesa ou perguntas em geral sobre a área de computação. O objetivo desse corpora é identificar padrões genéricos de conversação que serão utilizados para ampliar as possibilidades de formulação de perguntas sobre um determinado domínio do conhecimento.
2. **Corpora de *keywords*:** composto por textos que contenham palavras-chaves sobre um determinado assunto. O objetivo é obter padrões que poderão auxiliar na obtenção das respostas necessárias aos corpora de perguntas específicas.
3. **Corpora de perguntas específicas:** composto por textos que contenham perguntas específicas sobre um determinado domínio do conhecimento. O objetivo desses corpora é oferecer um "*know-how*" de perguntas sobre um domínio específico. Dessa forma, o chatterbot será capaz de responder uma série de perguntas que já foram feitas por outras pessoas.

3.3 Categorização e tratamento dos corpora utilizando o método LSI

Durante o processo de criação dos padrões, ocorrem muitas similaridades de natureza sintática e semântica, quando é feita a recuperação de dados a partir de textos com características de discussão, presentes em documentos extraídos de fóruns. Para evitar o tratamento de dados repetidos e facilitar a etapa da análise dos corpora pelos especialistas é necessário categorizar os padrões repetidos ou similares.

Mesmo em arquivos pequenos, a categorização manual seria impraticável e, no caso do problema proposto, o trabalho seria gigantesco. Seriam necessários vários meses de trabalho para categorizar manualmente os padrões de conversação. Para evitar a categorização manual, é proposta a categorização automática, através de técnicas de classificação automática de documentos baseadas no modelo vetorial[9], mais precisamente, nas técnicas baseadas no método *Latent Semantic Indexing* LSI.

No modelo vetorial, os documentos e consultas são representados por vetores no espaço euclidiano. Cada elemento do vetor de termos é considerado uma coordenada dimensional, na qual t é o número de termos e o peso W é dado pela posição do documento em cada dimensão. O peso do termo é o elemento que qualifica a relação entre o termo e o documento, além de especificar o tamanho e a direção do vetor que representa o documento. Os vetores no espaço euclidiano serão representados por uma

[9] Modelo de recuperação de informação no qual documentos e consultas são representados por vetores no espaço euclidiano de t-dimensões [Baeza-yates 1999].

matriz $\vec{M} = (M_{tN})$, a matriz de termos e documentos, sendo que as linhas representam os termos e as colunas representam os documentos. Cada elemento M_{tN} da matriz representa o peso de um termo em um determinado documento.

Antes de iniciar o processo de categorização dos padrões de conversão é necessário indexar os corpora. A indexação consiste na contagem das ocorrências das palavras nos seus respectivos arquivos. No problema específico da indexação de arquivos contendo padrões de conversação, cada linha do corpus será equivalente a um arquivo ou coluna na matriz $\vec{M}$. Além disso, nenhuma palavra será ignorada no processo de indexação, pois, para o problema em questão, isso iria prejudicar o processo de categorização.

Foi desenvolvido um programa na linguagem C ANSI especificamente para indexar os corpora e gerar uma matriz $\vec{M}$, sendo que as linhas são representadas pelos termos ou palavras e as colunas identificam o padrão no qual o termo está presente e os elementos identificam a freqüência do termo no padrão.

Para categorizar os padrões dos corpora será aplicado o algoritmo LSI na matriz $\vec{M}$. É necessário informar o percentual da matriz $\vec{M}$ que será utilizada e o percentual de similaridade a ser adotado no momento da categorização. Através do método *Singular Value Decomposition*[10] (SVD) é feita a decomposição da matriz $\vec{M}$ e, em seguida, é aplicado um fator de redução. Através da observação dos resultados gerados com diversas combinações, foi possível constatar que o valor mais indicado é 80% de utilização dos dados da matriz $\vec{M}$ e 75% de similaridade entre os padrões de conversação. A justificativa dos valores encontrados é devido à especificidade do problema de classificação dos padrões de conversação. Devido ao fato de cada padrão ter apenas 16 palavras, um descarte maior de dados da matriz $\vec{M}$ ou alteração no grau de similaridade poderia interferir nos resultados, ocasionando uma categorização indevida. Apesar da diferença sintática ser pequena, ela é expressiva em um padrão com 16 palavras. Apesar da equivalência na semântica, existe uma diferença de 4 palavras entre eles. Isso representa uma diferença de 25% no tamanho do padrão. Caso o grau de similaridade utilizado fosse maior, o algoritmo não seria capaz de agrupar determinados padrões. Um exemplo disso é apresentado no quadro 2.

topico descricao responsavel outras arquiteturas linux rodando em sparc alpha powerpc etc humberto sun com linux ?
topico descricao responsavel outras arquiteturas linux rodando em sparc alpha powerpc etc humberto windows no linux ?

Quadro 2: Padrões equivalentes com diferença semântica.

A diferença entre os dois padrões categorizados é de apenas duas palavras, um deles finaliza com “sun com linux?” e o outro “windows no linux?”. Esses padrões são idênticos semanticamente e devem fazer parte da mesma categoria.

[10] Está relacionado a técnicas matemáticas e estatísticas que incluem decomposição de auto-vetores e analises espectral e de fatores.

3.4 Manipulação dos arquivos pelos especialistas no assunto

Após a categorização, os especialistas no assunto farão uma varredura completa em todas as categorias geradas pelo LSI, com o objetivo de extrair apenas o que interessa para a criação das bases de conhecimento. Dos padrões presentes em uma determinada categoria, será extraído apenas um padrão dentre um conjunto de padrões idênticos na sintaxe e semântica. Os padrões resultantes desse processo irão formar os corpora descritos na sessão 3.3.

Ao executar a indexação e o algoritmo LSI, após a manipulação dos especialistas nos corpora de perguntas específicas junto com os de *keywords* e, em seguida, nos corpora de perguntas genéricas com os de *keywords*, teremos um conjunto de categorias que tendem a agrupar padrões de perguntas e suas respectivas respostas.

Ao final do processo, serão gerados diversos arquivos contendo duplas de padrões, sendo que o primeiro padrão corresponde à pergunta e o segundo, a resposta. O quadro 3 apresenta um exemplo dos arquivos contendo as duplas de padrões.

Como configurar o sun com o Linux? padrão com uma resposta aproximada
Como configurar um computador alpha para rodar Windows com Linux? padrão com uma resposta aproximada

Quadro 3. Dupla de padrões geradas pelos especialistas

4. Tuxbot – um estudo de caso na utilização de Chatterbot em ambientes de aprendizagem

Com o objetivo de facilitar a busca de informações contidas no site Focalinux, principalmente nas situações em que o visitante tem uma dúvida específica e não deseja ou não tem tempo de ficar pesquisando no site até achar a solução para o seu problema, é proposto o desenvolvimento do TuxBot. [Focalinux 2005].

TuxBot é um Chatterbot que responde perguntas sobre dúvidas de usuários do Linux. TuxBot utiliza a tecnologia ALICE através da implementação em Hypertext Preprocessor (PHP) com banco de dados Mysql. O conhecimento do TuxBot foi baseado inicialmente no guia Foca GNU/Linux e em padrões de conversação obtidos de sites na Internet. O Tuxbot está disponível no endereço http://www.ensino.org.br/tuxbot/.

Como estudo de caso, foram selecionados aproximadamente 9Mb de arquivos textos sobre o Linux, extraídos do site http://www.rau-to.unicamp.br[11]. Após a execução do programa em busca do texto “linux ?” o sistema gerou um corpus com 200 padrões. Ao término da execução do algoritmo LSI, foram geradas 6 categorias. Foi possível constatar que o LSI foi capaz de categorizar 100% dos padrões que deveriam ser categorizados na análise de um especialista no assunto.

[11] Sistema coorporativo de perguntas e respostas.

5. Considerações finais

O uso de chatterbots no apoio a alunos de ambientes virtuais de aprendizagem merece destaque e estudos mais profundos. O grande desafio dos pesquisadores da área está no desenvolvimento de soluções para a criação de bases de conhecimento que permitam ao chatterbot atender as expectativas e necessidades dos aprendizes.

Existem diversas soluções disponíveis para o uso ou desenvolvimento de chatterbots. A tecnologia ALICE é uma solução interessante e flexível, entretanto, de nada adianta um chatterbot sem uma boa base de conhecimento.

Através dos experimentos realizados foi constatado que a proposta de construção das bases de conhecimento aqui apresentada é viável e requer uma interferência pequena dos especialistas no processo de construção das bases de conhecimento em AIML.

Os resultados obtidos nas conversas com o Tuxbot demonstram que é possível criar uma base de conhecimento que possa atender às expectativas dos aprendizes, diminuir a sensação de isolamento, e aumentar a motivação.

6. Referências Bibliográficas

Alice, Alice Artificial Intelligence Foundation. Disponível em: http://www.alicebot.org> Acesso em: 01 mar. 2005.

Baeza-yates, R.; Ribeiro-neto, B. Modern Information Retrieval. Addison Wesley, 1999.

Deerwester, S; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Harshman, R. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6): 391{407), 1990. disponível em: http://citeseer.nj.nec.com/deerwester90indexing.

Firth, J. R. Papers in linguistics – 1934-1951. Oxford, Oxford University Press, 1957.

Focalinux, Disponível em: http://focalinux.cipsga.org.br Acesso em: 06.mar. 2005.

Partington, A. Patterns and meanings: using corpora for English language research and teaching. Amsterdã/Filadélfia, John Benjamins, 1998.

Sardinha, T. S. Lingüística de Corpus. Barueri, SP: Manole, 2004.

Teixeira, S.; Menezes, C. S. Facilitando o uso de Ambientes Virtuais através de Agentes de Conversação. XIV Simpósio Brasileiro de Informática na Educação - SBIE - 2003, Rio de Janeiro, RJ, Brasil, p. 483-492, 12 de Novembro de 2003.

Weizenbaum, J. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. Communications of the ACM Volume 9, Number 1 (January 1966): 36-45. Disponível em: http://i5.nyu.edu/~mm64/x52.9265/january1966.html Acesso em: 24.fev. 2003.