Uma Abordagem Semiautomática para Expansão e Enriquecimento Linguístico de Bases AIML para Chatbots

Sílvia M. W. Moraes
Faculdade de Informática - PUCRS
Av,Ipiranga, 6681 Prédio 32
Porto Alegre, RS, Brasil
+55 51 3320-3558
silvia.moraes@pucrs.br

Luciano Severo de Souza
Faculdade de Informática - PUCRS
Av,Ipiranga, 6681 Prédio 32
Porto Alegre, RS, Brasil
+55 51 3320-3558
luciano.severo@acad.pucrs.br

ABSTRACT
This paper describes a work in progress that proposes a semi-automatic approach to expand and enrich bases AIML of chatbots. The proposed approach consists of extracting textual information from FAQs to automatically expand the knowledge in a chatbot. The integration of the new base with existing bases requires manual refinement. The approach also allows enriching the current base with morphosyntactic information. This information aims to improve the matching between user input and patterns of existing questions in the base.

RESUMO
Este artigo descreve um trabalho em andamento que propõe uma abordagem semiautomática para expandir e enriquecer bases AIML de chatbots. A abordagem proposta consiste em extrair informações textuais de FAQs para expandir a base de conhecimento do chatbot. A integração da base assim definida com as existentes exige um refinamento manual. A abordagem permite ainda enriquecer a base atual do chatbot com informações morfossintáticas. Essas informações têm como objetivo de melhorar o casamento entre as entradas do usuário e os padrões de perguntas existentes na base.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing. K.3.2 [**Computers and Education**]: Computer and Information Sciencie Education – *Information systems education* .

General Terms
Experimentation.

Keywords
Chatbot, Natural Language Processing, AIML.

1. INTRODUÇÃO

A evolução da Internet aliada ao incremento, cada vez maior, de sistemas para *web*, especialmente para dispositivos móveis, tem provocado novas demandas em diferentes áreas do conhecimento. Segundo Banchs e Li [1], esse é o caso dos chatbots, os quais têm ganho muita popularidade e estão sendo considerados, atualmente, "canais modernos de comunicação entre empresas e clientes" [2]. Conceitualmente, os chatbots são sistemas de diálogo que têm como objetivo a comunicação, em linguagem natural, com usuários a fim de auxiliá-los de alguma forma [3,4]. Os chatbots, também chamados chatterbots ou agentes conversacionais, respondem a perguntas de tal forma que os usuários têm a impressão de estarem conversando com uma pessoa e não com um programa de computador [2].

Na área educacional, os chatbots têm recebido muita atenção. As razões de tal interesse estão centradas nas suas características. Os chatbots provêem uma interface mais natural e prática ao aluno. São capazes de oferecer um suporte pessoal, reconhecendo os pontos fortes, os interesses, bem como as habilidades individuais dos alunos. Além disso, podem acelerar o processo de aprendizagem ao atuarem como instigadores dos tópicos em estudo, resultando em um maior engajamento e na independência dos estudantes [5,6]. Os estudantes demonstram maior motivação quando há uma interação mais pessoal. Segundo Ghose e Barua [7], os alunos costumam preferir a interação com chatbots, mesmo quando a informação procurada está disponível na *web*.

Apesar da popularidade desses sistemas, a maioria dos chatterbots utiliza apenas ferramentas triviais da área de processamento da lingua natural (PLN). Além disso, suas bases de conhecimento, em geral, são escritas em AIML[1], que é baseada em casamento de padrões. Embora uma base de conhecimento AIML seja fácil de ser definida por exigir pouco conhecimento computacional para sua criação, precisam conter uma grande quantidade de padrões para que o chatbot matenha um diálogo próximo ao natural. Isso acaba aumentando, consideravelmente, o esforço manual necessário à sua construção, tornando o sistema pouco flexível para ser usado em contextos diferentes daquele para o qual foi projetado.

Como nosso objetivo é construir um chatterbot educacional, neste artigo propomos uma abordagem semiautomática para expandir (incluindo mais conhecimento) e enriquecer (acrescentando informações linguísticas) a base AIML de um agente conversacional. A expansão do conhecimento é realizada a partir da extração de informações textuais de bases tipo *Frequently Asked Questions* (FAQs). A ferramenta *Visual Interactive Syntax Learning (VISL[2])* é utilizada para prover as informações morfossintáticas à base AIML. Essas informações são usadas para melhorar o casamento entre as sentenças de entrada do usuário com os padrões de perguntas existentes na base.

Este artigo está organizado em 6 Seções. A Seção 1 faz uma breve introdução a chatterbots. A Seção 2 comenta alguns trabalhos relacionados a este. A Seção 3 faz uma breve introdução à AIML. A Seção 4 descreve o chatbot para o qual a abordagem proposta

[1] AIML significa Artificial Intelligence Markup Language. AIML é uma versão XML que foi projetada para criar sistemas de diálogo do tipo estímulo-resposta.

[2] A ferramenta VISL provê etiquetagem *on line* (http://beta.visl.sdu.dk/visl/pt/). Utiliza o parser PALAVRAS desenvolvido por Eckhard Bick [13].

foi usada. A Seção 5 descreve a abordagem usada. E, a Seção 6, apresenta, por fim, as nossas considerações finais e trabalhos futuros.

2. CHATTERBOTS

O termo chatterbot é uma junção das palavras *chatter* (pessoa que conversa) e *bot* (abreviatura de *robot* – robô). Embora o primeiro chatterbot tenha surgido na década de 60, tal termo foi usado pela primeira vez em 1994 por Michael Mauldin para designar um jogador controlado por computador [8]. Historicamente, o primeiro chatbot foi o ELIZA. Esse sistema foi construído em 1966 por Joseph Weizenbaum para simular um psicoterapeuta [9]. Da década de 60 para a atualidade, diferentes chatterbots foram criados. Sendo chatbot ALICE (Artificial Linguistic Internet Computer Entity) o mais conhecido. ALICE foi criada por Richard Wallace e foi ativada em 1995 [10]. Foi o primeiro programa com personalidade baseado em AIML.

Recentemente, os chatbots têm sido utilizados em diferentes domínios. Na área educacional, eles costumam ser usados em sistemas de tutoria, de perguntas e respostas, de conversação destinada a aprendizagem de uma nova língua, de diálogo para estimular a reflexão e as habilidades metacognitivas dos alunos. Atuam também como agentes pedagógicos e parceiros no processo de aprendizagem [5].

Apesar da popularidade desses sistemas, a maioria dos chatterbots utiliza apenas ferramentas triviais da área de processamento da lingua natural (PLN). Além disso, grande parte desses sistemas é baseada em casamento de padrões. Os chatterbots procuram por palavras-chave nas entradas do usuário e tentam associá-las a padrões existentes em uma base textual local, geralmente, definida em AIML [4,11]. Cada padrão possui ao menos uma resposta associada.

De acordo com Klüwer [11], a opção pela construção de chatterbots baseados em AIML se deve a duas razões: simplicidade no desenvolvimento e robutez no que se refere a entradas inesperadas. Apesar desses fatores, esse tipo de abordagem, além de gerar problemas quanto ao entendimento correto da entrada do usuário, exige que a base contenha uma grande quantidade de padrões. Isso aumenta consideravelmente o esforço manual necessário à construção e acaba tornando o sistema pouco flexível. Para contornar desses problemas, Klüwer sugere otimizações que vão desde a integração do sistema com outras bases de conhecimento ao uso de técnicas mais elaboradas de processamento da língua natural. E essas são as principais motivações para o nosso estudo.

Na seção seguinte, comentamos alguns trabalhos relacionados ao nosso

3. TRABALHOS RELACIONADOS

Existem na literatura diferentes abordagens que podem ser usadas para construção de bases AIML a partir de texto, no entanto poucas se referem à lingua portuguesa. Abu Shawar em [11] tenta demonstrar que a base de conhecimento de um *chatbot* pode ser definida a partir de qualquer texto e em qualquer idioma. A autora argumenta que não existe qualquer restrição, seja quanto ao tamanho ou quanto ao conteúdo do *corpus*. Para comprovar tal princípio, Shawar investiga *corpora* transcritos a partir de gravações de diálogos em inglês e em línguas africanas. Ela utiliza, em seu estudo, o *British National Corpus* (BNC) que não é um *corpus* de diálogo, mas possui um volume de dados considerável. Ela usa, ainda, o *Qur'an*, o livro sagrado do Islã, que também não é um *corpus* de diálogo. Seus experimentos são apresentados na forma de chatbots. Diferente da abordagem apresentada neste artigo, a autora não faz uso de qualquer tratamento linguístico. Sua abordagem para construir a base AIML é totalmente estatística. Ela construiu um chatbot para cada *corpus* mencionado. Para o *corpus* de línguas africanas, os resultados não foram satisfatórios devido a dois problemas principais: o domínio era limitado e existiam mais de dois participantes nos diálogos catalogados. No caso do BNC, o esforço computacional foi considerável e resultou em mais de 1 milhão de categorias organizadas em 800 arquivos. A avaliação do chatbot para o *Qur'an* foi subjetiva em razão da natureza do *corpus.* A percepção dos usuários que interagiram com esse chatbot foi negativa, pois nem sempre os usuários encontravam respostas para suas perguntas. A autora ainda realizou uma comparação do chatbot com o motor de busca do *Google.* A avaliação ocorreu a partir de uma lista com 15 questões pré-determinadas e mais 5 questões de livre escolha do usuário. As questões eram todas em língua inglesa. A autora relata que, em geral, 2/3 dos usuários preferiram utilizar o *chatbot* ao motor de busca. De acordo com os usuários, enquanto o chatbot fornecia respostas diretas, o Google retornava apenas uma série de *links.*

Bada e Menezes em [12] propõem uma arquitetura para chatbot baseada em agentes. A proposta dos autores também envolve a construção de bases AIML a partir de *corpus* usando informações morfossintáticas. Na arquitetura de Bada e Menezes, há agentes responsáveis pelo tratamento linguístico das sentenças do *corpus* e pela geração do código AIML correspondente. Um agente chamado Guru gerencia um módulo, escrito em Prolog, que tem por finalidade manter a base AIML e uma ontologia descrita na forma de proposições lógicas. Quando o interpretador AIML não encontra uma resposta na base AIML, o agente Guru tenta responder à pergunta usando as proposições lógicas da ontologia. De acordo com os autores, aliar técnicas de processamento da linguagem natural à pesquisa por padrões morfossintáticos resultou em uma abordagem eficaz para a construção automática bases de conhecimento para chatterbots.

A abordagem proposta se aproxima dos trabalhos estudados visto que extrai informações textuais para criar automaticamente a base AIML do chatbot. Os padrões gerados são constituídos de termos normalizados e anotados com informações morfossintáticas. A diferença da nossa abordagem está no fato que utilizarmos a abordagem também para expandir uma base AIML já existente. Nossa abordagem propõe algumas formas de contornar problemas decorrentes da integração.

Nas próximas seções fazemos uma breve introdução à linguagem AIML e ao chatbot usado em nosso estudo.

4. AIML

AIML é uma linguagem de marcação para Inteligência Artificial, desenvolvida por Richard Wallace, cujo objetivo é definir bases de conhecimento léxicas para agentes conversacionais [10]. As bases AIML estão organizadas em categorias. Para cada categoria, define-se um padrão de entrada correspondente a uma pergunta (*pattern*), para o qual podem ser associadas uma ou mais respostas (*templates*). O vocabulário AIML consiste em palavras, espaços e os caracteres especiais, tais como "*", conhecidos como "curingas". A função dos curingas é substituir partes de strings,

com o propósito de deixar os padrões de entrada mais abrangentes. Para que o chatbot não responda às mesmas perguntas sempre da mesma forma, pode-se usar as etiquetas *<random>* e *<li>*. Essas etiquetas permitem selecionar aleatoriamente uma resposta dentre várias. A Figura. 1 apresenta um trecho de código AIML.

```
<aiml version="2.0" encoding="UTF-8">
<category>
<pattern> OLA * </pattern>
<template>
        <random>
         <li> Oi! Muito prazer!</li>
         <li> Olá, como vai você? </li>
         <li> Oi! Tudo bem ? </li>
        </random>
    </template>
</category>
</aiml>
```

Figura 1. Trecho de código AIML.

Para que a interação com o usuário aconteça, é necessário um interpretador para AIML. Neste trabalho, foi utilizado o interpretador AB[3] versão 6.26, escrito em java. Cabe mencionar que para fins de padronização e simplificação no processo de casamento de padrões, os textos das perguntas são escritos com letras maiúsculas.

5. CHATBOT RICKY

O chatbot Ricky foi construído por 33 alunos da turma de Inteligência Artificial (IA) do curso de Sistemas de Informação, durante o semestre 2015/01. Esse chatbot foi projetado e construído manualmente pelos alunos. O propósito desse agente era educacional. O próprio processo de construção do chatbot foi usado como método de aprendizagem. Os alunos tinham que buscar conceitos sobre IA para popular a base AIML do agente. Como a ideia também era que o agente, após a sua implementação, fosse utilizado por outros estudantes com fins educativos, era importante criar um perfil com o qual os alunos pudessem se identificar. A Tabela 1 mostra o perfil que os alunos da turma de IA definiram para o Ricky.

É importante mencionar que, durante a construção do agente, pouco tratamento linguístico foi incluído. Basicamente, foram retirados apenas caracteres especiais das sentenças do usuário e aplicadas algumas transformações aos termos. Para estas, os alunos criaram um arquivo de substituições que permitiu transformar abreviaturas e termos reduzidos, comumente empregados na *web* (o internetês, exemplo: "vc"," n", "q", …), para as suas formas ortográficas "tradicionais" ("você", "não", "que", ...). Ao final da prototipação, o agente conversacional totalizou 26 bases AIML com aproximadamente 1.450 categorias. A Figura 2 mostra um trecho da base que os alunos implementaram.

Tabela 1. Perfil do chatbot Ricky.

Preferências	Livro: Guia dos Mochileiros da Galáxia; Série de TV: Game of Thrones; Música: eletrônica; Esporte: Futebol (gosta, mas não fala muito desde a última copa) e Xadrez. Time: seleção brasileira; Carro: Google Self-driving Car; Ídolo: Alan Turing.
Temas	Sistemas de Informação, Área de TI, Universidade, Faculdade, Esportes, Inteligência Artificial (subáreas: agentes, chatbots e alguns algoritmos de busca heurística)
Comportamento	Bem humorado em geral; empolgado quando fala de suas preferências, exceto no caso do Futebol; não gosta de ser ofendido, mas responde a insultos de forma educada; é técnico quando fala sobre temas como TI.

Observando-se o segundo pattern da Figura 2, podemos perceber que nenhum tipo de normalização morfológica foi utilizada. Isso significa que se o usuário realizar qualquer pergunta ao agente usando o termo "sistema" e não "sistemas", o agente não recuperará a resposta adequada. A abordagem que propomos procura resolver este problema.

```
…
  <pattern>* DIFERENCA * FRACA * FORTE *</pattern>
  <template> Resumidamente? Bom, IA forte quer criar computadores que pensam, enquanto que a IA fraca quer apenas resolver problemas sem possuir uma real inteligência...
  </template>
…
<pattern>* SISTEMAS DE INFORMACAO</pattern>
 <template>
 <random>
    <li>Legal, eu adoro falar sobre esse assunto! Sistemas de Informação é um dos cursos de tecnologia da Faculdade. … Sou suspeito em indicar esse curso pra alguem, ele simplesmente me encanta!</li>
    <li>Sistemas de Informação, agora tá falando a minha língua! Olha esse curso é um dos mais legais dentre os que existem na área de tecnologia da Facin. … Caso eu não tenha sido muito claro, você pode dar uma conferida no site do curso e se divertir por lá...</li>
   </random>
</template>
…
```

Figura 2. Trecho da base AIML construída pelos alunos.

O chatbot Ricky foi avaliado de forma subjetiva. Participaram de sua avaliação 17 pessoas. É importante ressaltar que nenhuma delas participou da implementação desse agente. Foram aplicados questionários de pré-teste e pós-teste aos avaliadores, os quais tinha entre 20 e 49 anos e atuavam em diferentes áreas de conhecimento. Apenas dois valiadores já haviam interagido com um chatbot antes. A maioria não tinha conhecimento algum em agentes reativos, processamento de linguagem natural e inteligência artificial. Quanto ao conhecimento em computação, 44% tinha conhecimento básico, 50% conhecimento intermediário ou avançado e apenas 6% não tinha qualquer conhecimento na área. As pessoas passaram, em média, 9 minutos interagindo com o chatbot Ricky. Comentaram sobre sua personalidade divertida e bem humorada, bem como sua rapidez nas respostas.

No entanto, apontaram como pontos fracos a dificuldade do agente no entendimento das perguntas feitas pelos usuários e a grande incoerência das respostas. O chatbot gerava muitas respostas inadequadas. Apesar do baixo desempenho do Ricky, como mostra a Tabela 2, várias pessoas responderam que voltariam a interagir com o agente. A Tabela 2 resume os resultados obtidos para os critérios: facilidade de uso (o quão fácil foi a interação do agente com o usuário), naturalidade (o quão próximo suas respostas estavam às humanas), robustez (sua capacidade de gerar uma resposta para qualquer pergunta) e

[3] Mais informações em https://code.google.com/p/program-ab/

coerência (o quanto as respostas eram adequadas às perguntas do usuário). Considerando os conceitos Razoável e Excelente, o agente conseguiu ao menos 60% de aceitação.

Tabela 2. Desempenho do chatbot Ricky.

Critério	Péssimo	Razoável	Excelente
Facilidade de Uso	29%	47%	24%
Naturalidade	29%	65%	6%
Robustez	29%	65%	6%
Coerência	41%	53%	6%

A seguir, descrevemos a abordagem usada para aprimorar a base do chatbot.

6. ABORDAGEM PROPOSTA

A construção de uma base de conhecimento AIML para *chatterbot* exige muito esforço manual, visto que sua diversidade e riqueza são fundamentais para que o agente tenha um bom e natural diálogo com o usuário. Uma alternativa para minimizar tal esforço é a construção ou expansão automática de bases de conhecimento a partir de textos.

É desejável que o *corpus* usado seja de diálogo e, preferencialmente, entre dois locutores. Não existem muitos *corpora* de diálogo disponíveis, especialmente para a língua portuguesa. Por esta razão, os pesquisadores que têm usado fóruns como fonte de conhecimento para os *chatterbots*. No caso dos fóruns, o desafio não está apenas em preprocessar os textos para definir os pares "pergunta-resposta". É desafiador também encontrar a resposta mais apropriada e confiável a uma pergunta, visto que em um fórum várias respostas de diferentes usuários podem ser postadas. Outra alternativa ainda são as *FAQs*, nas quais os pares "pergunta-resposta" já estão definidos e possuem um formato semiestruturado. Em nossa abordagem usamos FAQs.

6.1 Expansão da base AIML a partir de FAQs

Neste trabalho, foi usada a FAQ disponível em http://www.medicinaatual.com.br. Escolhemos esta FAQ em função da riqueza de informações disponíveis e da existência de um padrão claramente definido na estrutura de suas páginas HTML. Essa estrutura possibilitou a extração das informações a partir da ferramenta de análise para páginas *web* (*HTML parsers*) JSOUP[4]. Através desta ferramenta foi possível extrair e manipular elementos, atributos e textos oriundos de páginas escritas em linguagem HTML. Enquanto a maioria dos *sites* pesquisados disponibilizavam apenas algumas dezenas de pares "pergunta-resposta", neste site, foram encontradas aproximadamente 10.000 destes pares, distribuídos em 221 categorias (classes de doenças).

Para extrair as perguntas e respostas existentes nas FAQs foi escrito um *parser* em Java. Em seguida, as perguntas foram anotadas com informações linguísticas, preprocessadas e transformadas em padrões AIML. A Figura 3 mostra as etapas implementadas para construir a base AIML.

Figura 3. Etapas de construção da base AIML a partir de FAQs.

Essas etapas são detalhadas nas seções a seguir.

[4] http://jsoup.org

6.1.1 Anotação morfossintática

Após a extração realizada com o auxílio da biblioteca JSOUP foram obtidas aproximadamente 10.000 perguntas. Essas perguntas foram, então, anotadas pela ferramenta VISL para facilitar a extração dos seus elementos relevantes. Esse anotador segmenta a sentença em *tokens*[5] e usa lematização como forma de normalização morfológica. Na lematização, as palavras são transformadas em seus lemas. Substantivos e adjetivos são levados para forma masculino singular, e verbos, para o infinito. Por exemplo, as palavras "conectado" e "conectou", após o processo de lematização, seriam representadas por uma única forma, o lema "conectar".

O VISL anota *Part-Of-Speech* (POS), ou seja, atribui etiquetas que definem a classe gramatical das palavras. Os substantivos são rotulado como N, artigos como DET, verbos como V, preposições como PRP, etc. Esse anotador provê ainda diferentes tipos de anotação sintática as quais permitem identificar, por exemplo, o sujeito (@SUBJ), o verbo principal (@FMV) de uma sentença. Escolheu-se esta ferramenta pela qualidade da sua anotação para Língua Portuguesa, por ser bem conhecida e estar disponível para *web*. A anotação utilizada foi a "*Full morphosyntactic parse*". Para exemplificar, a anotação provida pela ferramenta, considere a sentença "Quais são os principais sintomas da asma?". A etiquetagem resultante desta sentença-exemplo pode ser visualizada na Figura 4.

quais [qual] <interr> **DET** M/F P @SC>
são [ser] <fmc> **V** PR 3P IND VFIN @FMV
os [o] <artd> **DET** M P @>N
principais [principal] <SUP> **ADJ** M P @>N
sintomas [sintoma] <sick-c> **N** F P @<SUBJ
de [de] <sam-> **PRP** @N<
a [o] <-sam> <artd> **DET** F S @>N
asma [asma] <sick> **N** F S @P<

Figura 4. Anotação da ferramenta VISL.

6.1.2 Extração de padrões

Após a etiquetagem das perguntas, a etapa seguinte consistiu em extrair das mesmas os tokens mais relevantes. Esses tokens são usados para compor os *patterns* AIML. O intuito desta etapa é otimizar o casamento de padrões, eliminando palavras desnecessárias que poderiam influenciar de forma negativa o desempenho do chatbot. A ausência de um *token* significativo pode levar a um casamento incorreto, gerando consequentemente uma resposta inadequada. Ao passo que um *token* irrelevante pode inviabilizar totalmente um correto casamento de padrões, o que poderá resultar na ativação de uma resposta universal[6]. O primeiro passo, então, foi definir o formato dos padrões que seriam considerados. A exemplo de Bada e Menezes em [9], o formato de padrão escolhido foi baseado em pronomes interrogativos, sujeito, verbo principal e seus complementos. De acordo com aqueles autores, esses elementos tornam viável a construção de uma base de conhecimento AIML consistente e possível de ser usada por chatbots.

[5] Os *tokens*, em geral, são palavras. A ferramenta VISL reconhece como tokens alguns termos compostos tal como: Presidente da República.

[6] Entende-se por resposta universal aquelas que são apresentadas quando não houve casamento entre a pergunta do usuário e os *patterns* previstos na base de conhecimento do chatbot. São respostas de contingência.

A Figura 5 mostra o formato principal de padrão que foi usado neste estudo e sua aplicação na sentença-exemplo: "Quais são os principais sintomas da asma?". Cabe mencionar que o ponto-e-vírgula é meramente um separador dos termos e que o último substantivo (N) corresponde ao primeiro complemento encontrado. Além disso, nos padrões foram usados apenas os lemas dos termos existentes nas perguntas. Todas as perguntas da base foram formatadas como mostrado. Se uma pergunta, por ventura, não tivesse algum dos elementos considerados, o padrão era gerado apenas com os existentes.

Formato: *<interr>; V* @FMV*; N @SUBJ>; N*

Exemplo: *[qual]<interr>;* [ser] **V** @FMV; [sintoma] **N** @<SUBJ; [asma] **N** @P<

Figura 5. Formato dos padrões.

6.1.3 Conversão para AIML

Definidos os formatos dos padrões, a próxima etapa foi a conversão para AIML. Ao transformar as perguntas em *patterns* AIML, duas transformações foram aplicadas para facilitar o casamento de padrões: definição de subpadrões e uso do curinga "*". Os subpadrões são resultantes da combinação de ao menos 3 termos do padrão original. Após uma análise manual de uma amostra da FAQ anotada, foi definido que padrões com 3 termos seriam um bom caso de estudo. Essa quantidade tornaria o padrão mais genérico a fim de aumentar as chances de encontrar um padrão compatível com a pergunta do usuário e, ao mesmo tempo, específico para evitar casamentos incorretos ou ativação de universais. Optou-se também por manter, nos subpadrões, os substantivos. Na amostra analisada, os substantivos expressavam, em geral, a semântica da sentença. A Figura 6 exibe os subpadrões que foram incluídos na base, o primeiro subpadrão foi definido sem o verbo [ser]. Já o segundo, sem o pronome interrogativo [qual].

Padrão Original : *[qual]<interr>;* [ser] **V** @FMV; [sintoma] **N** @<SUBJ; [asma] **N** @P<

Subpadrão 1 : *[qual]<interr>;* [sintoma] **N** @<SUBJ; [asma] **N** @P<

Subpadrão 2 : [ser] **V** @FMV; [sintoma] **N** @<SUBJ; [asma] **N** @P<

Figura 6. Padrão original e subpadrões.

Cabe mencionar ainda que nenhum tratamento é aplicado às respostas. Elas são simplesmente agregadas, na íntegra, à cláusula *<template>*. A Figura 7 mostra um trecho da base AIML gerada. Os *tokens* extraídos das perguntas da FAQ são convertidos para caixa alta e intercalados pelo curinga "*". Foi necessário também retirar a acentuação para reduzir a quantidade de variações de um mesmo padrão. A acentuação, na língua portuguesa, é, sem dúvida, uma forma de diferenciar os termos. No entanto, como tal remoção foi realizada após o processo de anotação linguística e cada padrão faz uso de ao menos 3 termos, acredita-se que os termos em conjunto conseguem determinar o contexto semântico correto da pergunta.

A base AIML do chatbot foi organizada por doença. Sendo assim, foi gerado um arquivo AIML para cada doença. Essas bases também foram avaliadas. Usamos o mesmo processo avaliativo aplicado ao chatbot Ricky. A avaliação foi realizada por 5 usuários que responderam os questionários de pré-teste e pós-teste. Participaram desse processo avaliativo 3 usuários da área de Enfermagem, 1 da área de Adminstração e 1 de Tecnologia da Informação (TI). Excetuando o usuário da área de TI, todos os demais nunca haviam interagindo com um chatbot antes e não conheciam as tecnologias envolvidas em sua construção.

*<pattern> QUAL * SINTOMA * ASMA </pattern>*
<template>
Dispneia, tosse e sibilância torácica constituem a tríade clássica de sintomas associados à asma. No entanto, ...
</template>

Figura 7. Trecho da base AIML gerada automaticamente.

Cada pessoa interagiu por cerca de 20 minutos com chatbot cuja base foi gerada a partir das FAQs. A Tabela 3 mostra a avaliação feita por esses usuários. Analisando-se os resultados do critério "facilidade", pode-se inferir que a interface, embora simples e puramente textual não prejudicou a interação com o usuário. A "naturalidade" teve um desempenho bom, provavelmente porque as respostas da FAQ (elaboradas manualmente pelos autores da FAQ) foram mantidas na íntegra. A ausência de tratamento de sinônimos, como esperado, no entanto, teve impacto nos resultados desse critério. A "robustez" também teve um resultado interessante, indicando que a construção de subpadrões é um caminho viável. Já o critério "coerência" não foi satisfatório. O principal problema neste caso foi a ausência de variantes da mesma sentença. Isso é uma limitação de abordagens baseadas em *corpus*, pois só se consegue extrair os termos que estão de fato no texto. Os usuários faziam perguntas para as quais havia resposta na FAQ, no entanto o chatbot não as achava pois a estrutura linguística era diferente da representada na base.

Foi solicitado ainda que o usuário apontasse pontos fracos e fortes do chatbot. Como pontos fortes, eles indicaram rapidez, simplicidade e praticidade. Na opinião dos usuários, é mais simples, mais fácil e mais rápido interagir com um chatbot do que consultar a FAQ diretamente. O principal ponto fraco apontado foi a falta de coerência e não tratamento adequado de palavras-chave.

Tabela 3. Análise do desempenho das bases AIML extraídas automaticamente das FAQs.

Critério	Péssimo	Razoável	Excelente
Facilidade	0	40%	60%
Naturalidade	20%	40%	40%
Robustez	20%	60%	20%
Coerência	40%	40%	20%

É importante frisar que os avaliadores do chatbot Ricky e da base AIML gerada a partir de FAQ foram diferentes. Apesar do número menor de avaliadores desta última, observamos que os resultados, proporcionamente, foram muito próximos. Em ambas avaliações, a soma dos índices razoável e excelente se aproximou de 80%. Apenas no caso da coerência, a soma desses índices não passou de 60%. Esperamos com a integração das bases, melhorar o desempenho do chatbot, especialmente, no aspecto coerência

6.2 Integração das bases AIML

Atualmente, estamos implementando a integração das bases AIML do chatbot Ricky, que foram elaboradas manualmente, com as bases AIML, geradas automaticamente. As bases do chatbot Ricky tem padrões bastante heterogêneos e não possuem qualquer tratamento linguístico. Nosso primeiro desafio foi identificar a

variedade dos padrões existentes na base AIML do Ricky. Foram encontrados padrões redundantes, incorretos e em formatos diversos. A Figura 8 exibe alguns dos formatos encontrados.

(1) VOCE CONHECE A HISTORIA DA IA
(2) CASA PREFERIDA DE GAME OF THRONES
(3) QUAL SUA CASA PREFERIDA DE GAME OF THRONES
(4) CASA QUE VOCE ACHA LEGAL DE GAME OF THRONES
(5) FALA SOBRE O *
(6)* SERIE GUERRA DOS TRONOS

Figura 8. Exemplo de padrões com formatos heterogêneos.

Nem todos os alunos usaram os recursos do AIML (como *). Alguns esqueceram, durante a edição, das transformações. Por exemplo, a abreviatura IA, usado no *pattern* (1) será sempre transformada em "Inteligência Artificial", logo esse padrão não será encontrado pelo interpretador AIML. Os *patterns* (2), (3) e (4) semanticamente são idênticos. Várias bases não têm a mesma preocupação com os sinônimos. Embora sinônimos sejam necessários, incluí-los diretamente na base não é a melhor estratégia, pois deixará a base muito extensa. O *pattern* (6) também se refere a GAME OF THRONES, mas usa o título em português. O *pattern* (5) utiliza recursos AIML mas é muito genérico, o que o torna muito similar a vários outros padrões. Para resolver os problemas identificados definimos um *pipeline* (Figura 9) para viabilizar a integração das bases.

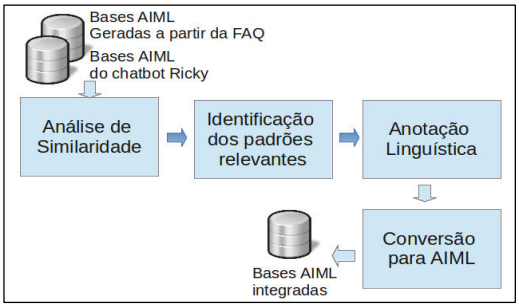

Figura 9. Etapas de integração das bases.

Na análise de similaridade estamos usando medidas usuais de identificação similaridade entre strings para detectar padrões semelhantes. Comparamos os padrões das bases do Ricky com os gerados automáticamente pela FAQ. O maior índice que coincidência que estamos encontrando, no entanto, está na própria base do Ricky. Ainda estamos testando as medidas de similaridade e índices mais adequados. Após a análise similaridade, o sistema faz uma proposta de unificação para os padrões considerados semelhantes. A unificação exige supervisão humana. Os padrões unificados são submetidos à anotação linguística, a qual é feita pela ferramenta VISL. Feito isso, novos padrões AIML são, então, gerados. Após a integração, vamos novamente avaliar o chatbot junto aos usuários.

7. CONSIDERAÇÕES FINAIS

Apresentamos nesse artigo uma proposta de expansão usando FAQ e enriquecimento de bases AIML com informações linguísticas. Os resultados ainda são preliminares, mas animadores. Nossos próximos passos, incluem, além de uma nova avaliação do chatbot, o tratamento de sinônimos, a extração de informações de outras FAQs e de outros *corpora*, e, ainda, o uso métodos recomendação de conteúdo para indicação de materiais de estudo.

8. AGRADECIMENTO

Esta pesquisa tem o apoio financeiro da Pontifícia Universidade Católica do Rio Grande do Sul (EDITAL N. 02/2014 – Chamada para o Programa de Apoio à Atuação de Professores Horistas em Atividades de Pesquisa na PUCRS).

9. REFERÊNCIAS

[1] Banchs, Rafael E. and Li, Haizhou. "IRIS: a Chat-oriented Dialogue System based on the Vector Space Model". Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic, Korea, p. 37-42 (2012).

[2] Kuligowska, K. and Lasek, M. "Virtual assistants support customer relations and business processes". Information Management, Gdańsk University Press (2011).

[3] Jurafsky, Daniel e Martin, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2005).

[4] Al-Zubaide, Hadeel e Issa, Ayman A. "OnBot: Ontology based ChatBot". Fourth International Symposium on Innovation in Information & Communication Technology, IEEE, p. 7-12 (2011).

[5] Kerly, Alice; Ellis, Richard e Bull, Susan. "Conversational Agents in E-Learning". Applications and Innovations in Intelligent Systems XVI, Springer, p. 169-182 (2009).

[6] Griol, David; Garcia-Herreno, Jesús e Molina, José M. "The EduAgent Platform: Intelligent Conversational Agents for E-learning Applications". Ambient Intelligence – Sofware and Applications. Advances in Intelligent and Soft Computing, Volume 92, Springer, p. 117-124 (2011).

[7] Glose, Supratip e Barua, Jagat Joyti. "Toward the implementation of a Topic specific Dialogue based Natural Language Chatbot as an Undergraduate Advisor", IEEE (2013).

[8] Mauldin, Michael L. "Chatterbots, TinyMuds, and the Turing Test Entering the Loebner Prize Competition, AAAI'94, Proceedings..., p. 16-21 (1994).

[9] Weizenbaum, Joseph. "ELIZA – a computer program for the study of natural language communication between man and machine",ACM, 9(1), New York, USA, p. 36-45 (1996).

[10] Wallace, Richard. "The Anatomy of ALICE", http://www.alicebot.org, (2003).

[11] Shawar, A. A Corpus Based Approach to Generalising a Chatbot System. Procesamieno del Lenguaje Natural, 31, (2003).

[12] Bada, E; Menezes, C. Uma proposta para extração de perguntas e respostas de textos, TISE, vol.8,(2012).

[13] Bick, E. The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework (2000