Proceedings of Symposium in Information and Human Language Technology. Uberlândia, MG, Brazil, October 2–5, 2017.

Geração de perguntas e respostas para a base de conhecimento de um chatterbot educacional

Joyce Martins, Camila V. Martins

Departamento de Sistemas e Computação
Universidade Regional de Blumenau (FURB) – Blumenau, SC – Brazil
joyce@furb.br, camila.viviani@outlook.com

Abstract. *This paper presents a chatterbot that answers or asks questions in Portuguese about a basic education text. From an input text composed of simple sentences, the questions and answers are generated in the chatterbot knowledge base. For this purpose, a morphosyntactic analysis is made and the semantic roles of each word are obtained. Seven semantic roles are treated, from which is defined what pronoun or adverb should be used to elaborate the question. Although with limitations, due to the complexity of Portuguese, it is possible to generate questions and answers from any text composed of simple sentences, and not only from those of basic education.*

Resumo. *Este artigo apresenta um chatterbot que responde ou faz perguntas em língua portuguesa sobre um texto da educação básica. A partir de um texto de entrada composto por sentenças simples, são geradas as perguntas e as respostas da base de conhecimento do chatterbot. Para tanto, efetua-se a análise morfossintática e obtêm-se os papéis semânticos de cada palavra que compõe o texto. São tratados sete papéis semânticos, a partir dos quais define-se o pronome ou o advérbio interrogativo que deve ser usado para elaborar a pergunta. Embora com limitações devido à complexidade da língua portuguesa, é possível gerar perguntas e respostas para qualquer texto composto por sentenças simples, e não apenas para os da educação básica.*

1. Introdução

Estudos mostram que o uso de perguntas/respostas no processo de ensino-aprendizagem é benéfico [Chi et al., 1994 apud Le; Kojiri e Pinkwart, 2011]. Os autores afirmam que fazer perguntas específicas ajuda a identificar a falta de conhecimento sobre determinado assunto. Descrevem também aplicações educacionais que, com o uso de perguntas e respostas geradas automaticamente, têm por objetivo: a aquisição de habilidades, a avaliação do conhecimento ou o diálogo com tutores.

Dentre as diversas ferramentas computacionais disponíveis, é possível encontrar os *chatterbots*. Segundo Comarella e Café [2008], um *chatterbot* é um software que simula uma conversação com um ser humano, proporcionando para o usuário, no caso, o estudante, uma experiência semelhante a que teria numa conversa on-line com um especialista de uma determinada área. De maneira simplista, um *chatterbot* possui uma interface para entrada das perguntas (ou mensagens) formuladas pelo usuário. Em seguida, procura a resposta correspondente na sua base de conhecimento, que, se for encontrada, será apresentada ao usuário. Caso contrário, uma resposta padrão do tipo

Desculpe, não entendi o que você falou! será emitida. Souza e Moraes [2015] afirmam que é necessário despender um tempo considerável para criar manualmente uma base de conhecimento de um *chatterbot*. Afirmam ainda que "Por esta razão, têm surgido abordagens que procuram gerar automaticamente as bases desses agentes a partir de *corpora* existentes, inclusive a partir de informações disponíveis na *web*" [Souza e Moraes, 2015].

Assim sendo, este trabalho também apresenta uma proposta para gerar automaticamente perguntas e respostas da base de conhecimento de um *chatterbot* educacional a partir de um texto da educação básica. Nas seções seguintes a arquitetura do *chatterbot* e o processamento do texto de entrada são descritos, bem como as considerações finais do trabalho são apresentadas.

2. Arquitetura do ChatterEDU

O processamento realizado pelo *chatterbot*, chamado de ChatterEDU, se dá em etapas. Primeiro o usuário deve entrar com um texto com conhecimentos da educação básica. O texto de entrada passa por um processamento onde são obtidos o papel semântico[1] e a classificação morfossintática de cada palavra, de forma similar à abordagem proposta por Amancio, Duran e Aluisio [2011]. Em seguida, a partir das sentenças de entrada são geradas perguntas e respostas, as quais são gravadas em uma base de conhecimento Artificial Intelligence Markup Language[2] (AIML), de forma a ser possível realizar uma conversação em linguagem natural o mais próximo possível de uma conversa entre seres humanos.

Finalizado o processamento do texto, o usuário é direcionado para a interface de conversação com o ChatterEDU. O usuário pode iniciar a conversação com uma mensagem de saudação ou diretamente com uma pergunta relacionada ao texto inserido. A mensagem (ou a pergunta) é enviada para o *chatterbot*, que utiliza o interpretador Program AB [Wallace, 2013] para buscar a resposta correspondente nas bases de conhecimento AIML previamente criadas. Se for encontrada uma resposta correspondente à mensagem (ou pergunta), a mesma será retornada, caso contrário, será informado que o *chatterbot* não possui conhecimento sobre a mensagem (ou pergunta) inserida. O ChatterEDU possui três bases de conhecimento: uma com saudações, outra com respostas a serem usadas quando o *chatterbot* não consegue responder às perguntas do usuário e a terceira com perguntas e respostas sobre o texto de entrada. As duas primeiras foram criadas manualmente enquanto a última é gerada automaticamente.

3. Processamento do Texto de Entrada

A complexidade da aplicação desenvolvida está no processamento do texto de entrada para gerar automaticamente as perguntas e as respostas e, em seguida, criar a base de

[1] Kipper [2005 apud Scarton, 2013, p. xv] diz que os papéis semânticos "descrevem a relação semântica subjacente entre um verbo (ou predicador) e seus argumentos e são usados para descrever padrões léxicos e semânticos no comportamento dos verbos."

[2] A AIML é uma linguagem de marcação baseada na eXtensible Markup Language (XML), usada para especificar as bases de conhecimentos dos *chatterbots* [Wallace, 2013]. Algumas *tags* básicas da linguagem são: `<category>`, agrupa perguntas e respostas; `<pattern>`, define uma possível pergunta ou mensagem; `<template>`, indica a resposta correspondente a um determinado `<pattern>`.

conhecimento AIML. Primeiramente, o texto de entrada é dividido em sentenças, que são enviadas para o *parser* Palavras[3] [Bick, 2000]. Nesse processamento é utilizada a opção *semantic roles*, que determina o papel semântico de cada palavra. Para exemplificar, tem-se a frase *Blumenau sofreu uma grande enchente em 2008*. Toma-se a análise do *parser* para a palavra *Blumenau* (Quadro 1). Destacam-se as seguintes informações: (a) classe gramatical: `PROP` (nome próprio); (b) flexões: `M/F` (masculino/feminino), `S/P` (singular/plural); (c) função sintática: `@SUBJ>` (sujeito); (d) papel semântico: `§AG` (agente).

```
Blumenau [Blumenau] PROP M/F S/P @SUBJ> §AG #1->2
sofreu [sofrer] <fmc> V PS 3S IND VFIN @FS-STA §PRED #2->0
(...)
em [em] PRP @<ADVL #6->2
2008 [2008] <card> NUM M/F P @P< §LOC-TMP #7->6
```

Quadro 1. Análise morfossintática

A segunda etapa do processamento consiste em verificar qual o papel semântico de cada palavra para definir a pergunta que será feita. Os pronomes e os advérbios interrogativos usados para cada papel semântico podem ser vistos no Quadro 2. Para delimitar o escopo do projeto, dentre os papéis semânticos identificados pelo *parser* Palavras, foram selecionados os mais comuns em textos da área de conhecimento Geografia. Assim, os papéis semânticos tratados pelo ChatterEDU são: (a) AG: agente; (b) LOC: lugar; (c) LOC-TMP: localização temporal (dia, mês, ano, indicação de tempo); (d) ORI-TMP: origem temporal (dia, mês, ano, indicação de tempo); (e) EXT: extensão ou quantidade; (f) EXT-TMP: período de tempo; (g) TH: tema.

papel semântico	pronome / advérbio	frase de entrada	pergunta formulada
AG – voz ativa	Quem	Os barrigas-verdes moram em Santa Catarina.	Quem mora em Santa Catarina?
AG – voz passiva	Por quem	Santa Catarina é habitada por barrigas-verdes.	Santa Catarina é habitada por quem?
LOC	Onde	Santa Catarina fica na região sul.	Onde Santa Catarina fica?
LOC-TMP	Quando	Blumenau sofreu uma grande enchente em 2008.	Quando Blumenau sofreu uma grande enchente?
ORI-TMP	Desde quando	Desde 1852 foram registradas 64 enchentes em Blumenau.	Desde quando foram registradas 64 enchentes em Blumenau?
EXT	Quanto	O estado mede 95703 quilômetros quadrado.	O estado mede quanto?
EXT-TMP	Quanto tempo	A tragédia durou por duas semanas.	A tragédia durou quanto tempo?
TH	O que	Florianópolis tem cerca de 421 mil habitantes.	O que tem cerca de 421 mil habitantes?
TH – verbo ser	Qual	Blumenau é a terceira maior cidade de Santa Catarina.	Qual é a terceira maior cidade de Santa Catarina?
TH – humano	Quem	Ele foi empregado em diversas missões.	Quem foi empregado em diversas missões?

Quadro 2. Papéis semânticos e pronomes / advérbios interrogativos

No Quadro 2 é possível notar que para alguns papéis semânticos pode-se elaborar mais de um tipo de pergunta, dependendo de outros termos linguísticos

[3] Disponível em <http://visl.sdu.dk/visl/pt/parsing/automatic/parse.php>.

presentes no texto de entrada. No caso do papel semântico AG, por exemplo, é verificado se o verbo está na voz ativa ou passiva, para então definir se será utilizado *quem* ou *por quem*. Além disso, é feito um tratamento em casos específicos para gerar perguntas com conjugação verbal correta. Todos os verbos regulares são tratados, assim como alguns verbos irregulares (ser, ir, estar e dar), no presente, pretérito perfeito, pretérito imperfeito e pretérito-mais-que-perfeito.

Independente do papel semântico, são geradas três tipos de perguntas: a pergunta que o usuário pode fazer ao ChatterEDU, o tema sobre o qual o usuário deseja que o *chatterbot* faça perguntas e a pergunta do *chatterbot* sobre um determinado tema. Para elucidar como são geradas as perguntas e respostas, considera-se a frase do Quadro 1, na qual foram identificados dois dos papéis semânticos tratados: AG e LOC-TMP. Nesse caso, são geradas as seguintes perguntas: (a) pergunta do usuário (AG) - *^quem^sofreu^ enchente^2008^*; (b) pergunta do usuário (LOC-TMP) - *Quando^Blumenau sofreu^ enchente^*; (c) tema (AG) - *^sobre^quem^ sofreu^enchente^2008^*; (d) tema (LOC-TMP) - *^sobre^quando Blumenau sofreu^ enchente^*; (e) pergunta do *chatterbot* (AG) - *Quem sofreu uma grande enchente em 2008?*; (f) pergunta do *chatterbot* (LOC-TMP) - *Quando Blumenau sofreu uma grande enchente?*

Observa-se que tanto para a pergunta do usuário quanto para o tema da pergunta, é usado o caracter ^, que em AIML representa um ponto da sentença que pode ou não possuir mais de uma palavra qualquer, permitindo uma maior gama de perguntas reconhecidas. Por exemplo, se a sentença *Eu quero saber quem sofreu uma enchente em 2008*. estiver na base de conhecimento AIML, o usuário tem que entrar exatamente com essa frase para ser possível encontrar a resposta correspondente. Mas, se a sentença gravada for *^quem^sofreu^enchente^2008^*, o usuário pode entrar com uma variação da pergunta, tal como: *Eu gostaria de saber quem sofreu uma grande enchente em 2008*. ou *Quem sofreu uma enchente no ano de 2008?*.

Para o tema da pergunta utiliza-se a mesma ideia, mas com a palavra *sobre* na frente. Então, quando o usuário quiser responder a perguntas feitas pelo ChatterEDU, deve entrar com uma frase tal como *Quero falar sobre quem sofreu enchente em 2008*. Nesse caso, o ChatterEDU faz a pergunta (do *chatterbot*) referente ao assunto solicitado pelo usuário. Já a pergunta do *chatterbot* é a mais simples de ser elaborada. Isto porque não há necessidade de substituir nenhuma palavra por curinga, uma vez que a pergunta que o ChatterEDU deve fazer para o usuário deve estar completamente gravada na base de conhecimento. Desta forma, é adicionado o pronome ou o advérbio interrogativo no início ou no final da frase, excluindo da sentença a resposta desejada, como exemplificado no Quadro 2.

Quanto às respostas, são geradas duas: uma resposta curta e outra completa (a própria sentença de entrada). Também são geradas respostas padrões para indicar para o usuário que ele acertou ou errou ao responder uma pergunta feita pelo ChatterEDU. Por fim, é possível entrar com a frase *Faça perguntas sobre o texto*., onde as palavras *sobre* e *texto* são obrigatórias. Nesse caso, o ChatterEDU fará aleatoriamente uma das perguntas existentes na sua base de conhecimento e verificará se a resposta informada está ou não correta. O último passo realizado é gravar a base de conhecimento AIML. Cabe ressaltar que cada vez que ocorrer esse processamento, a base de conhecimento

anteriormente criada será sobrescrita. Então, se o usuário inserir um texto e logo após outro, o ChatterEDU manterá uma conversação apenas sobre o último texto inserido.

4. Considerações Finais

Esse trabalho descreveu o desenvolvimento do ChatterEDU, uma aplicação desktop que interage com o usuário em língua portuguesa. A partir do processamento de um texto, em princípio, da educação básica na área de conhecimento Geografia, gera automaticamente, com base nos papéis semânticos das palavras que o compõem, perguntas e respostas da base de conhecimento, possibilitando que o usuário faça ou responda perguntas sobre o texto. Um texto para ser processado deve estar gramaticalmente correto e ser composto apenas por sentenças simples.

Apesar da limitação quanto ao tipo de sentença e papéis semânticos tratados, o uso desses permitiu gerar perguntas e respostas de textos de qualquer área de conhecimento. Mas, mesmo com as restrições estabelecidas, existem sentenças que não são processadas adequadamente, gerando perguntas inconsistentes (incompletas ou contendo parte da resposta), com erros ortográficos ou erros de concordância verbal. Por fim, observou-se demora em processar os textos de entrada, já que é necessário acesso on-line ao *parser* Palavras para efetuar a análise morfossintática e determinar os papéis semânticos. Para processar uma frase com sete palavras, a aplicação leva em média 6s (segundos), já um texto com quarenta e uma palavras, leva em torno de 15s. Além disso, a linguagem AIML tem suas próprias limitações que forçam o usuário a fazer perguntas e fornecer respostas utilizando os termos da base de conhecimento. Por conta disso, mesmo que o usuário responda corretamente uma pergunta, o *chatterbot* pode retornar erro se a resposta não estiver exatamente como gravada na base AIML.

Referências

Amancio, M. A.; Duran, M. S.; Aluísio, S. M. (2011), Automatic question categorization: a new approach for text elaboration. Procesamiento del Lenguaje Natural, v. 46, p. 43-50, 2011.

Bick, E. (2000), The parsing system PALAVRAS: automatic gramatical analysis of Portuguese in a constraint grammar framework. Aarhus University Press.

Comarella, R. L. e Café, L. M. A. (2008), “Chatterbot: conceito, características, tipologia e construção”, Informação & Sociedade: estudos, João Pessoa, v. 18, n. 2, p. 55-67, maio/ago. 2008.

Le, NT.; Kojiri, T. e Pinkwart, N. (2011), “Automatic question generation for educational applications: the state of art”. In: van Do, T.; Thi, H. e Nguyen, N. (eds) Advanced computational methods for knowledge engineering. Springer, Berlin.

Scarton, C. (2013), VerbNet.Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. Mestrado. Instituto de Ciências Matemáticas e de Computação, USP.

Souza, L. S. e Moraes, S. M. W. (2015), “Construção automática de uma base AIML para chatbot: um estudo baseado na extração de informações a partir de FAQs”. In *Anais do XII ENIAC*. Natal, RN. p. 137-141.

Wallace, R. S. (2013), “ALICE A.I. Foundation”. http://alicebot.blogspot.com.br/.