PROCESSAMENTO DE LINGUAGEM NATURAL, ROBÔS DE CONVERSAÇÃO E LINGUÍSTICA

Edio Roberto Manfio
prof.ediorobertomanfio@gmail.com
Faculdade de Tecnologia de Garça/UEL

Abstract

This article aims to present the relationship between Natural language processing and Linguistics, as well as an example of application of PLN in a conversational robot that can be used on a daily basis for various types of users. The study was done from queries in books and articles about the issues involved and the application of that knowledge in a system developed for this purpose: the robot named Professor Tical. The research demonstrates that the interrelation between different spheres of knowledge of both the humanities and the exact area is inevitable and, since always, quite productive.

Resumo

O presente artigo tem por objetivo apresentar a relação que existe entre Processamento de Linguagem Natural e Linguística, além de um exemplo de aplicação do PLN em um robô de conversação que pode ser utilizado no dia a dia por vários tipos de usuários. O estudo foi feito a partir de consultas em livros e artigos sobre os assuntos envolvidos e da aplicação desses conhecimentos em um sistema desenvolvido para esse fim: o robô denominado Professor Tical. A pesquisa demonstra que a inter-relação entre diferentes esferas de conhecimento tanto da área de humanas quanto da área de exatas é inevitável e, desde sempre, bastante produtiva.

1 Introdução

O Processamento de Linguagem Natural (doravante apenas PLN) é uma das áreas de conhecimento que compõem as bases da Inteligência Artificial. Normalmente, quando se fala em PLN quase sempre se está tratando do processamento de linguagem escrita, mormente fazendo o uso de conhecimentos léxicos, sintáticos e semânticos. Estão, portanto, incluídos no conceito de PLN a compreensão, geração e outras tarefas como tradução idiomas, cuja precisão depende em muito do processamento de linguagens naturais (RICH, 1993; SCHILDT, 1989).

Em outras palavras, muito do que se aprende na escola durante o ensino fundamental e médio em termos de linguagem é utilizado pelo PLN quando de sua

aplicação em sistemas autônomos. Evidentemente que, para a efetiva aplicação em sistemas, os critérios são outros e teorias linguísticas mais complexas também estão envolvidas no processo.

Entre muitos exemplos de sistemas que operam com o conceito de PLN, podem-se citar os *chatbots*. Nesse trabalho, não só há menção sobre PLN e sua relação com a Linguística como também há à disposição um exemplo de aplicação dessa área de conhecimento em algo utilizável no dia a dia.

2 Robôs de Conversação e PLN

Diferentemente dos chats convencionais, os *chatbots* - ou robôs de conversação - são máquinas com as quais os humanos interagem por meio da linguagem escrita. O que há de diferente entre esses e os chats (*Conversational Hypertext Access Technology*) convencionais, em que deve haver pessoas interagindo com pessoas por meio do computador, é que os *chatbots* são programados para tentar imitar um interlocutor em conversação. Em relação a um interlocutor humano, eles têm o diferencial de trabalhar com um banco de dados bem mais preciso em se tratando de um assunto específico, ou seja, não corre o risco de esquecer ou confundir alguma informação.

Para esse estudo, será dada preferência à expressão 'robô de conversação' - como também são conhecidos - ou simplesmente 'robô', embora eventualmente seja possível fazer referência também aos anglicismos *chatbot* ou *chatterbots*, mesmo que não sejam sinônimos perfeitos.

Entre os robôs de conversação mais antigos, um dos mais famosos e historicamente conhecidos é o Eliza, idealizado por Joseph Weizenbaum em 1966. Programado para comportar-se como um psicanalista, apresentou resultados bastante interessantes na qualidade de experimento pois, durante o período em que foi testado, muitas pessoas pensaram estar de fato falando com um analista.

Tendo o Eliza com uma das referências, verifica-se que é perfeitamente possível criar um robô que possa operar com o PLN, não necessariamente para simular um analista, mas com outras funções que possam atender a necessidades diversas. Uma aplicação do PLN utilizada aqui como exemplo é o Professor Tical, criado para responder a perguntas sobre Dialetologia, Geossociolinguística, ALiB - Atlas Linguístico do Brasil (PROJETO, 2014; COMITÊ; 2001) e teorias afins.

O robô opera partindo dos conceitos básicos de PLN e faz também o uso de tabelas tipo *hash*. Como é de se prever em tecnologias desse tipo, em versões vindouras e mais complexas, é possível dotar o sistema com recursos de Processamento de Sinais da Fala - PSF – mais especificamente *síntese* e *reconhecimento de voz*[1].

[1] Dois estudos que relacionam essas duas áreas do Processamento de Sinais da Fala com Dialetologia e Geossociolinguística são "Aspectos fonéticos no *DosVox* enquanto aplicativo tipo texto-fala" (MANFIO, 2014a) e "Como funcionam alguns fonemas no aplicativo *Balabolka*" (MANFIO, 2014b).

Como se sabe, coexistem diversos modos como uma mesma palavra é pronunciada no Brasil e a variedade lexical que dizem respeito a um mesmo conceito pode ser explorada paralelamente tanto pelo PLN quanto pelo PSF. Em PLN isso pode ser feito acrescentando sintaxes e vocabulário. Em PSF, aprimorando vozes sintéticas e reconhecimento (MANFIO, 2014c). No presente momento, entretanto, será dada atenção tão somente com PLN e tabelas *hash*.

O *hash* é basicamente uma estrutura de dados específica com o poder de associar palavras-chave a elementos correlacionados. As tabelas *hash* Também são conhecidas como ‘tabelas de dispersão’ ou ‘tabelas de espalhamento’, e possuem como característica diferencial fazer buscas rápidas conseguindo valor correlacionado a partir de uma entrada simples, ou seja, uma palavra ou pequena expressão. O armazenamento físico dos dados durante a execução do Professor Tical é feito através de arquivos XML (eXtensible Markup Language). Importante lembrar que em se tratando de tabelas *hash*, ‘palavras-chave’ podem ser chamadas de ‘chaves de pesquisa’ o os ‘elementos correlacionados’ também denominam-se ‘valores’.

Implementado na linguagem C++ utilizando o Visual Studio 2010 da Microsoft, esse protótipo funciona na plataforma Windows 32 e 64 bits, considerando-se os sistemas operacionais XP, Seven e Windows 8. O banco de dados disposto é composto por perguntas e respostas bastante básicas e em um número bastante reduzido pois tem inicialmente apenas a função atender aos experimentos iniciais e ao conteúdo desse estudo.

3 Linguística e PLN

Como dito anteriormente, há uma relação bastante estreita entre a Linguística e o PLN. Interessantemente, ambas as áreas de conhecimento fazem parte de um conjunto de saberes necessários aos estudos da Inteligência Artificial, que por sua vez também é uma área de conhecimento. O PLN lança mão normalmente da Gramática Gerativo Transformacional (doravante apenas GGT) que pertence a um ramo da Linguística Geral. Além disso, muito do que já foi desenvolvido em termos de softwares para PLN fez o uso das linguagens de programação PROLOG, LISP e C. Com exceção dessa última, PROLOG e LISP baseiam-se na GGT (ARARIBÓIA, 1988; RICH, 1993; SCHILDT, 1989). Embora pareça confuso, é possível visualizar as relações e intersecções entre essas esferas de conhecimento de modo bastante simples como o representado no esquema a seguir:

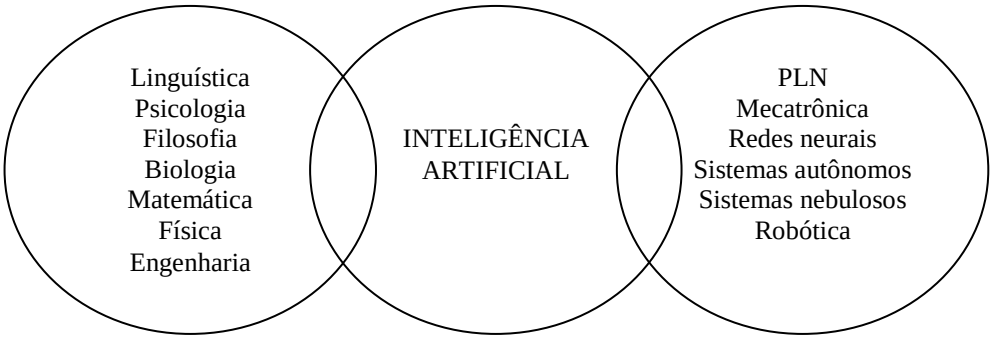

Figura 01 – Diferentes áreas de conhecimento e suas inter-relações

Assim representado, pode-se notar que, embora pareça a muitos pesquisadores de áreas específicas estarem muito distantes disciplinas da área de humanas como a Linguística e outras da área de exatas como a linguagem PROLOG.

Importante lembrar que a GGT encontra-se incluída na Linguística e, especificamente sobre ela, vale lembrar que trata-se de um tipo de gramática que deu contribuições importantes para o ensino de línguas. Criada por Noam Chomsky em meados da década de 1950, tinha o diferencial de destacar a criatividade quanto à utilização da língua, diferentemente do estruturalismo que se detinha mormente à estrutura da linguagem. De acordo com a teoria, a capacidade de distinguir sentenças gramaticais de não-gramaticais e o potencial de produzir e compreender um número infinito de sentenças gramaticais era o que realmente tinha importância. Uma frase como "O computador grita", embora seja gramaticalmente correta, torna-se problemática em nossa cultura, pois se não estiver devidamente contextualizada, pode não ter sentido, diferentemente de "O computador calcula", que não gera estranhamento. Note-se, no entanto, que ambas possuem sujeito verbo e predicado, concordâncias verbo-nominais e ortografia corretas.

Um sistema autônomo de PLN que esteja programado para avaliar sintaxes deve estar preparado para isso, ou seja, deve considerar como inadequada ou inaceitável a frase "O computador grita".

4 Considerações sobre o robô

O nome do Professor Tical é uma sigla que foi criada a partir de uma expressão que encerra muito resumidamente área envolvida e de todas as teorias a ela associadas. Observando dessa forma e mesmo considerando que há uma infinidade de robôs de conversação funcionando na grande rede, um que responda sobre tais temas e em língua portuguesa certamente é novidade.

A área de conhecimento temática escolhida para o Professor Tical é relevante por vários motivos: é a cada dia maior a necessidade de buscar manter o idioma ativo na Sociedade da Informação e o protótipo tem potencial de contribuir para a proliferação de interfaces de acesso a banco de dados baseadas em PLN; pode-se dizer que praticamente não há robôs de conversação que versem sobre Linguística e, muitos menos, em português; Tical pode operar, no mínimo, como agente de difusão cultural da área de conhecimento sobre as quais ele responde.

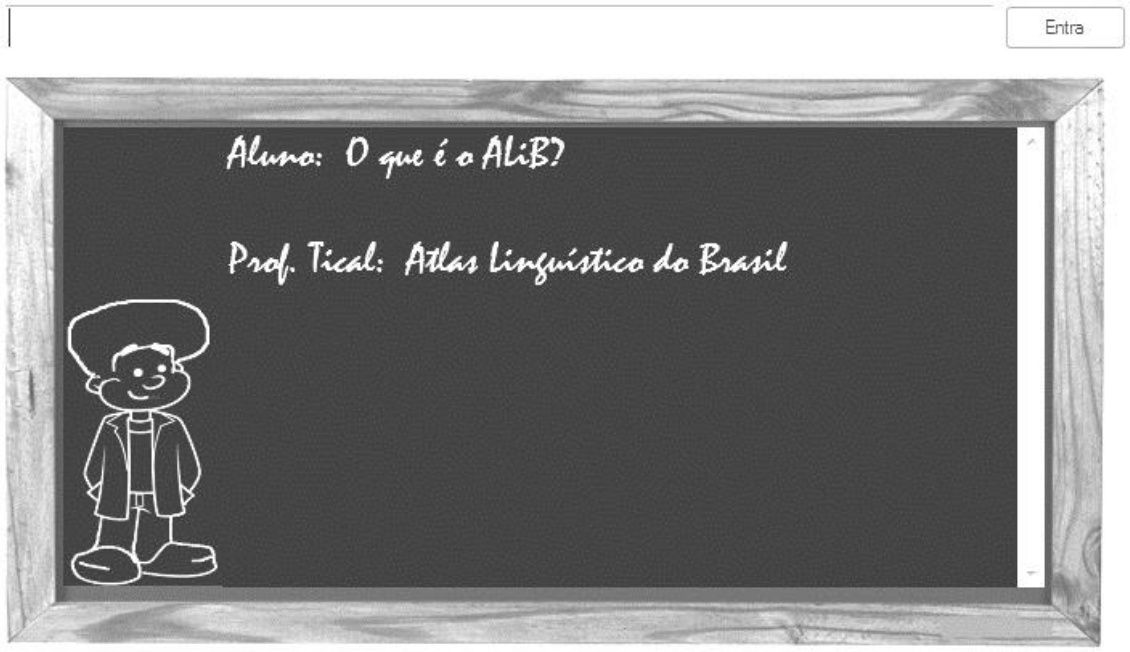

Figura 02 – Impressão de tela inteira do Professor Tical

A Figura 2 evidencia como esse protótipo apresenta notória simplicidade em sua aparência e funcionalidade. O ambiente é significativamente amigável e os elementos que vão ao encontro disso são o personagem à esquerda que representa o Professor Tical e um fundo simulando uma lousa, que sugere interação em uma sala de aula. Nota-se também que nos diálogos há as indicações 'Professor Tical' e 'aluno' que distinguem os respectivos turnos dos interlocutores.

Normalmente, todas as medidas tomadas por desenvolvedores no que diz respeito aos processos de personalização e caracterização de elementos temáticos pode parecer um tanto desnecessárias ou irrelevantes a alguns pesquisadores, porém, são procedimentos necessários quando se trata de aprimorar quesitos como ambiência amigável e funcionalidade.

As necessidades de se criar robôs que versem sobre os mais diversos tipos de assuntos tem também justificativas das mais diversas. Interessante salientar que os robôs de conversação costumam atender quase às mesmas funções básicas da maior parte dos jogos. Dentre essas funções básicas podem-se destacar as mais relevantes que são entretenimento, treinamento, educação e difusão cultural (HUIZINGA, 2007).

O Robô Ed (CONPET, 2014), disponibilizado na página do Conpet - Programa Nacional da Racionalização do uso dos derivados do petróleo e do gás natural é um exemplo entre muitos. Ele foi idealizado para fornecer informações sobre assuntos ligados ao meio ambiente mas não deixa de ter um viés de entretenimento e treinamento - para quem está disposto a estudar, avaliar e promover o conteúdo disponibilizado.

Mesmo o clássico Eliza, embora criado experimentalmente para aplicação de algumas teorias e tecnologias da época, ainda hoje continua sendo utilizado como objeto de estudo e/ou para diversão. Válido lembrar que ele também impulsiona de modo indireto

algumas discussões sobre psicanálise: o robô ‘ouve pacientemente’ o que a pessoa tem a dizer e a deixa à vontade para falar sobre o que quiser.

Relevante salientar que o Professor Tical não foi criado para portar-se como um psicanalista, porém tem mais ou menos a mesma funcionalidade básica do Eliza considerando-se as condições de produção e complexidade envolvidas. Em outras palavras, ambos - Tical e Eliza – tem a função de aplicar tipos de tecnologias e respectivas teorias e, ao mesmo tempo, impulsionam uma certa esfera de conhecimento: Eliza evidencia o PLN de modo geral; Tical, a Dialetologia, Geossociolinguística e ALiB.

Desse modo, a expectativa é que Tical, enquanto aplicativo, seja utilizado por diversas pessoas do modo como melhor lhes convier uma vez que uma das vantagens sistema que é a interagir com o robô de forma dialógica ao invés de fazer a escolha entre de lista ou palavras-chaves sublinhadas em um parágrafo (ROTHERMEL, 2007).

5 Considerações finais

A proposta de deste estudo foi apresentar a relação direta entre PLN e Linguística, além de um exemplo de aplicação do PLN em um sistema que pode ser utilizado no dia a dia por vários tipos de usuários. Foi possível verificar que o PLN é utilizado em uma infinidade de aplicações práticas com funções bastante variadas, ou seja, desde simples entretenimento até pesquisa.

Entre os exemplos elencados, destacam-se os robôs de conversação em função de seu amplo potencial de aplicabilidade nas diferentes esferas de conhecimento. Especificamente sobre este tipo de aplicação do PLN, apresentou-se um protótipo de robô de conversação denominado Tical que foi criado para responder a perguntas sobre Dialetologia, Geossociolinguística, ALiB - Atlas Linguístico do Brasil.

Portanto, a interdisciplinaridade e inter-relação entre diferentes esferas de conhecimento tanto da área de humanas quanto da área de exatas é notória, inevitável e indissociável. Entre elas, o Processamento de Linguagem Natural, a Linguística e as linguagens de programação utilizadas na construção de robôs de conversação aqui apresentadas são apenas algumas das muitas que cooperam no desenvolvimento e/ou aprimoramento de tecnologias interativas.

Referências

ARARIBÓIA, G. *Inteligência Artificial*. Rio de Janeiro: Livros Técnicos e Científicos Editora Ltda., 1988.

COMITÊ Nacional do Projeto ALiB. *Atlas Linguístico do Brasil:* questionário 2001/Comitê Nacional do Projeto ALiB. - Londrina: Editora UEL, 2001.

CONPET - Programa Nacional da Racionalização do uso dos derivados do petróleo e do gás natural. Robô Ed. Disponível em: <http://www.ed.conpet.gov.br/br/converse.php>. Acesso em: 24 abr. 2014.

HUIZINGA, Johan. *Homo Ludens:* o jogo como elemento da cultura. 4 ed. São Paulo: Perspectiva, 2007.

MANFIO, Edio Roberto. *Aspectos fonéticos no DosVox enquanto aplicativo tipo texto-fala*. *In XXI Seminário do CELLIP* - Paranaguá, 2013. Disponível em: <http://cellip.files.wordpress.com/2013/08/anais-do-xxi-cellip.pdf>. Acesso em: 21 jun. 2014a.

MANFIO, Edio Roberto. Como funcionam alguns fonemas no aplicativo *Balabolka*. In *Revista Via Litterae*. Anápolis, v. 4, n. 2, p. 191-204, jul./dez. 2012. Disponível em: <www2.unucseh.ueg.br/vialitterae>. Acesso em: 21 jun. 2014b.

MANFIO, Edio Roberto. Processamento de Linguagem Natural, Processamento de Sinais da Fala, Geolinguística e um Naco de Humor. In *Anais do X Seminário de Iniciação Científica Estudos Linguísticos e Literários - Sóletras*. UENP – Jacarezinho, 2013. Disponível em: <http://www.cj.uenp.edu.br/index.php/institucional/eventos/1-soletras/event_details>. Acesso em: 21 jun. 2014c.

PROJETO Atlas Linguístico do Brasil. Disponível em: <http://twiki.ufba.br/twiki/bin/view/Alib/WebHome>. Acesso em: 02 mai. 2014.

RICH, Elaine. *Inteligência Artificial*. Tradução Maria Cláudia Santos Ribeiro Ratto. São Paulo: Makron Books, 1993.

ROTHERMEL Alessandra. *Maria*: Um *chatterbot* desenvolvido para os estudantes da disciplina "Métodos e Técnicas de Pesquisa em Administração". In: Simpósio de Excelência em Gestão e Tecnologia. Disponível em: <http://www.aedb.br/seget/artigos07/1429_artigos2007eget.pdf>. Acesso em: 21 jun. 2014.

SCHILDT, Herbert. *Inteligência Artificial utilizando linguagem C*. Tradução Cláudio Gaiger Silveira e Mônica Soares Rufino. São Paulo: McGraw-Hill, 1989.