Desenvolvimento de um sistema de pergunta e resposta baseado em corpus

Adriano Jorge Soares Arrigo
Universidade do Sagrado Coração, Bauru/SP
e-mail: arrigo.adriano@gmail.com

Elvio Gilberto Silva
Universidade do Sagrado Coração, Bauru/SP
e-mail: elvio.silva@usc.br

Henrique Pachioni Martins
Universidade do Sagrado Coração, Bauru/SP
e-mail: henrique.martins@usc.br

Patrick Pereira Silva
Universidade do Sagrado Coração, Bauru/SP
e-mail: patrick.pedreira@usc.br

Comunicação oral

Pesquisa concluída

Palavras Chave: Sistema de Pergunta e Resposta; Padrões de Superfície Textual; PLN.

INTRODUÇÃO

O desenvolvimento da capacidade humana de ler e escrever usando os computares como auxílio de estratégia de ensino mostrou que existe uma nova relação e um grande potencial na capacidade de aprendizagem do aluno com o resultado da interação com essas máquinas. (CANTALICE, 2002).

De acordo com o relatório Panorama setorial da Internet do Centro de Estudos das Tecnologias da Informação e da Comunicação no Brasil (CETIC.br) (2013) 69% dos alunos acessam a Internet diariamente. Porém, segundo a pesquisa “TIC Kids Online Brasil 2012”, apenas 42% das crianças e adolescentes que acessam a Internet declaram não saber diferenciar sites que possuem uma informação verdadeira ou não. (CETIC.BR, 2013). Assim, torna-se importante

encontrar métodos e desenvolver ferramentas que possam processar e organizar tais informações de modo automático, oferecendo meios de identificar a veracidade de uma resposta.

Dentre as várias ferramentas que, indiretamente, calculam a precisão e a veracidade uma informação podemos citar os Sistemas de Pergunta e Resposta (Question Answering System) ou simplesmente, Sistemas de Q&A. Tais sistemas aceleram e automatizam a obtenção da informação pretendida, eliminando a intervenção humana na filtragem dos documentos e na leitura de texto, devolvendo uma resposta sucinta retirada dos documentos que analisa.

Assim, diante desta problemática, novas metodologias e ferramentas que aprimorem o processo de busca de informações online, com foco particular do processamento da língua portuguesa, tornam-se relevantes.

METODOLOGIA

Para o processo da recuperação dos padrões de respostas foi necessário a construção de um corpus dividido em seis domínios de conhecimento, a saber: DESCOBRIDOR, FAMOSO-POR, ANO-NASCIMENTO, LOCAL, DEFINIÇÃO e INVENTOR. Foram feitas, pelo menos, duas perguntas distintas relacionadas a cada domínio. Por exemplo, para o domínio DEFINIÇÃO, foram feitas perguntas para definir alguma coisa, como “O que é uma Timbila?”. Para essa questão, o corpus é construído na recuperação de informação para todos os resultados da query “Timbila+ instrumento musical”, que são as palavras chaves para a pergunta e a resposta, etiquetadas no corpus como <PERGUNTA> e <RESPOSTA>, respectivamente.

Foram utilizados comandos de linguagens típicas para recuperação de padrões textuais como sed e awk (Dougherty and Robbins 1990), linhas de comando e alguns utilitários textuais, como grep (Bambenek and Klus 2009). Essas linguagens utilizam-se de expressões regulares para poder formar comandos capazes de recuperarem sentenças específicas em arquivos textos.

Foram construídos dois algoritmos: um para a aprendizagem de padrões de respostas e o outro para calcular a precisão desses padrões. O algoritmo 1 é responsável por copiar os 100 primeiros resultados da query formulada a partir da pergunta inserida no buscador web, considerando apenas o conteúdo textual das páginas Web, além de trazer essas informações previamente formatadas, ou seja, somente com o conteúdo textual. As sentenças que se mantiveram durante todo esse processo são ranqueadas para, ao final do algoritmo 1, sejam selecionados os cinco padrões de respostas mais frequentes.

O algoritmo 2 analisa a precisão de cada padrão de resposta selecionado. Para isso, os padrões selecionados são submetidos a um mecanismo de busca para que seja verificada a sua capacidade de recuperação de informação (precisão).

No exemplo da pergunta “O que é uma Timbila?”, o melhor padrão de resposta, de acordo com os resultados do algoritmo 1, foi “<PERGUNTA> é o nome de um <RESPOSTA>”. Assim, esse padrão é inserido no buscador na forma “timbila é o nome de um”, omitindo a resposta para justamente recuperar as palavras que se encaixam nesse padrão. Dessa forma, são possíveis os seguintes resultados:

1 – Timbila é o nome de um instrumento musical. (resposta correta).

2 – Timbila é o nome de um <QUALQUER_PALAVRA>;

É calculada a precisão pela frequência absoluta acumulada, uma adaptação para o método descrito no trabalho de Ravichandran e Rovy (2002), que consiste em P = Ca/Co, onde:

Ca = total de sentenças recuperadas que contenha a resposta correta;

Co = total de sentenças recuperadas.

Dessa forma, é obtida a precisão de cada padrão de resposta.

RESULTADOS E DISCUSSÕES

Como forma de testar o sistema, foi elaborado um experimento que envolveu a avaliação dos padrões encontrados por meio do corpus. O experimento consistiu em

apresentar, para o mecanismo de busca Google, um conjunto de 12 perguntas, duas relacionadas a cada domínio específico (DESCOBRIMENTO, ANO NASCIMENTO, FAMOSO POR, DEFINIÇÃO, INVENTOR, LOCALIZAÇÃO). A tabela 1 mostra o resumo dos resultados encontrados.

Tabela 1 – Respostas das perguntas e padrões encontrados e suas respectivas precisões

Domínio	Pergunta	Padrão	Precisão
DESCOBRIDOR	Quem descobriu Netuno?	<PERGUNTA> foi descoberto por <RESPOSTA>	0.58
		<PERGUNTA> descobridor <RESPOSTA>	0.25
	Quem descobriu o Nióbio?	<PERGUNTA> descoberto pelo <RESPOSTA>	1
		<PERGUNTA> descoberto por <RESPOSTA>	1
		<PERGUNTA> descobridor <RESPOSTA>	0.95
ANO-NASCIMENTO	Em que ano Brad Pitt nasceu?	<PERGUNTA> nasceu em <RESPOSTA>	0.4
		<PERGUNTA> nascimento <RESPOSTA>	0.3
	Em que ano Gilberto Gil nasceu?	<PERGUNTA> nasceu no <RESPOSTA>	0.29
		<PERGUNTA> nasceu em <RESPOSTA>	0.24
FAMOSO-POR	Daniel Radcliffe é/ficou famoso pelo quê?	<PERGUNTA> famoso por <RESPOSTA>	1
		<PERGUNTA> famoso por interpretar <RESPOSTA>	0.42
	Courtney Cox é/ficou famosa pelo quê?	<PERGUNTA> famosa por interpretar <RESPOSTA>	0.47
		<PERGUNTA> famosa por <RESPOSTA>	0
DEFINIÇÃO	O que é Timbila?	<PERGUNTA> é o nome de um <RESPOSTA> é um <PERGUNTA>	0.83
		<PERGUNTA>, um <RESPOSTA>	0.75
	O que é concretismo?	<PERGUNTA> corresponde a um <RESPOSTA>	1
		<PERGUNTA> foi um <RESPOSTA-1>*	0.65
		<PERGUNTA> foi um <RESPOSTA-2>*	0.15
INVENTOR	Quem inventou o Pantógrafo?	<PERGUNTA> foi construído por <RESPOSTA>	0.66
		<PERGUNTA> foi inventado pelo <RESPOSTA>	0.58
		<PERGUNTA> inventor do <RESPOSTA>	0
	Quem inventou a lâmpada?	<RESPOSTA> inventa a <PERGUNTA>	1
		<RESPOSTA> inventor da <PERGUNTA>	0.47
		<RESPOSTA> inventou a <PERGUNTA>	0.29
LOCALIZAÇÃO	Onde fica a Islândia?	<PERGUNTA> situado no <RESPOSTA>	0.58
	Onde fica a Chapada da Diamantina?	<PERGUNTA> na <RESPOSTA>	0.26
		<PERGUNTA> - <RESPOSTA>	0

Fonte: Elaborada pelo autor (2014)

*Nesse caso, foram aceitas as respostas “movimento vanguardista” (<RESPOSTA-1>) e “movimento artístico” (<RESPOSTA 2>).

CONSIDERAÇÕES FINAIS

Apesar de incipientes, os resultados foram satisfatórios pois, de um modo geral, permitiram recuperar respostas pontuais para questionamentos realizados juntos aos motores de busca; o que pode indicar potencialidades na abordagem sugerida nesta investigação.

Entretanto, ajustes com relação a alguns padrões devem ser feitos já que, nem todos permitem recuperar uma resposta de modo tão preciso. Fatores como a própria complexidade da língua portuguesa dificultam essa tarefa. A grafia das palavras em português é muito rica. Isso pode ser visto em palavras como “*lampada”* e *“lâmpada”, “artistico”* e *“artístico”*, *“pantógrafo”* e *“pantografo”*. A web é muito extensa, e há, assim, uma pluralidade muito grande nas formas de escrever as palavras, embora, algumas vezes, a ortografia das palavras esteja incorreta.

Apesar das limitações da metodologia proposta, o foco particular do processamento da língua portuguesa, é a principal contribuição deste trabalho, não só para a validação de uma metodologia de extração de padrões na web, principalmente porque há muito poucos recursos para esta língua, uma vez que a maior parte das pesquisas envolve o idioma inglês.

Referências

BAMBENEK, J.; KLUS, A. **Grep Pocket Reference**. California: O'REILLY, 2009.

CANTALICE, L., M. de. **Tecnologia na educação**. Psicol. Esc. Educ. Campinas v. 6, n. 2, p. 187, Dec. 2002.

CETIC.BR. **O uso da Internet por alunos brasileiros de EnsinoFundamental e Médio**. Ano 5. Número 2. São Paulo, SP. 2013. 12f. Disponível em: <http://cetic.br/media/docs/publicacoes/6/panorama-setorial-agosto-2013.pdf>. Acesso em: 01 mar. 2015.

DOUGHERTY, D., ROBBINS, A. **Sed & Awk,** Second Edition. California: O'REILLY, 1990.

RAVICHANDRAN, D., HOVY, E. Learning Surface Text Patterns for a Question Answering System. **Proceedings…** 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002.