Sistema de perguntas e respostas com uso de informação morfossintática

Erick G. Maziero[1], Felipe Gomes[1]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Caixa Postal: 668 - CEP: 13560-970 – São Carlos – SP – Brasil
erickgm@grad.icmc.usp.br, felipe.gomes@gmail.com

Abstract. *This paper describes the development of a questions and answers system, which, given wh-questions questions, seeks the answer in a set of texts. The main feature of the system is the use of a part-of-speech tagger to obtain the words of interest (nouns and numbers) of both question and texts, used to search for the answer. This work is a superficial investigation of the challenge of responding to questions automatically.*

Resumo. *Este artigo descreve o desenvolvimento de um sistema de perguntas e respostas, que, dada uma pergunta do tipo "fechada", busca a resposta em um conjunto de textos. A principal característica do sistema é a utilização de um etiquetador morfossintático para a obtenção das palavras de interesse (substantivos e numerais) tanto da pergunta quanto dos textos, na busca pela resposta. Este trabalho trata-se de uma investigação superficial do desafio de responder a perguntas automaticamente.*

1. Introdução

Sistemas de resposta automática a perguntas é um grande desafio na área de Processamento de Língua Natural (PLN), em que, dada uma pergunta, esta deve ser interpretada e sua resposta é buscada em um conjunto de textos. Esses sistemas são assessorados por diversas outras ferramentas de PLN, como etiquetadores, corpora, dicionários e bases de dados lexicais e até sistemas automáticos de tradução [Gonzalo 2001]. São inúmeras as estratégias e abordagens desses tipos de sistemas, combinando ferramentas e técnicas, atuando sobre uma ou mais línguas.

O sistema de perguntas e respostas automático desenvolvido e descrito neste trabalho é uma abordagem superficial do problema, dado o primeiro contato dos autores com a problemática de responder a perguntas automaticamente.

2. Desenvolvimento

O sistema foi desenvolvido em linguagem Perl, dada sua versatilidade no processamento de textos e um etiquetador (MXPOST, [RATNAPARKHI 1996]) foi utilizado para a obtenção da classe das palavras presentes tanto nas perguntas quanto nos textos.

O sistema objetiva responder a uma pergunta, que é do tipo do tipo "fechado", sendo introduzida por palavras como "quem", "que", "quais" e afins. Este tipo de pergunta, em sua maioria, requererá como resposta um substantivo ou um numeral.

2.1. Arquitetura do Sistema

Motivado pelo tipo de pergunta a ser respondida (tipo “fechado”), o sistema realiza a leitura e etiquetação da pergunta a fim de obter os substantivos e numerais para posterior montagem de uma expressão regular para busca nos textos.

O etiquetador utilizado é o MXPOST, que utiliza o modelo probabilístico de Máxima Entropia para determinar a etiqueta de *Part-Of-Speech* (POS) de cada parte do texto (*token*). Esse etiquetador é disponível na Web e apresenta bons desempenhos em experimentos realizados tanto para o português quanto para o inglês.

Feita a etiquetação e a extração dos substantivos e numerais contidos na pergunta, uma simples expressão regular é composta. Esta é contida dos substantivos e numerais, que são intercalados com o padrão “.*”, que casa com qualquer conteúdo. Assim, considere a seguinte pergunta exemplo: *“Qual é o nome do estádio que sedia o amistoso entre Brasil e Itália em fevereiro de 2008?”* Feito sua etiquetação, os substantivos e numerais são identificados, extraídos e utilizados na montagem da seguinte expressão regular:

*nome.*estádio.*amistoso.*Brasil.*Itália.*fevereiro.*2008*

A expressão regular acima é utilizada na busca em todos os textos da coleção, que podem conter a resposta à pergunta. A busca é realizada sentença a sentença nos textos utilizando um operador de casamento de padrões da linguagem utilizada sem diferenciar letras maiúsculas ou minúsculas.

Essa expressão casará com qualquer sentença que contenha as palavras nome, estádio, amistoso, Brasil, Itália, fevereiro e 2008. Isto obriga o casamento com uma sentença que contenha todas as palavras listadas e na ordem em que aparecem na expressão regular. Quando ocorre um casamento da expressão regular montada com uma sentença, esta é etiquetada e seus substantivos e numerais são extraídos e armazenados a fim de comporem a resposta final à pergunta. Aqui é armazenado o nome do arquivo que contém o texto da sentença em que houve o casamento do padrão.

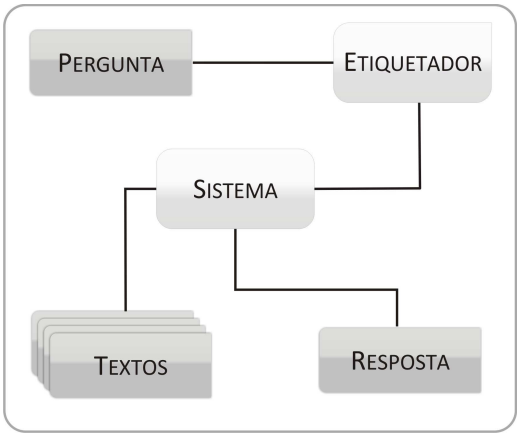

Figura 1. Arquitetura do Sistema

Após todos os textos terem sido processados, as respostas são montadas e gravadas em um arquivo seguindo o modelo “[substantivo1, substantivo2, numeral1, etc.;texto_x]”, onde os substantivos e numerais são listados dentro dos colchetes,

separados por vírgulas e, após um ponto e vírgula vem o nome do arquivo que contém o texto, de onde foram extraídas as palavras.

A Figura 1 mostra a arquitetura do sistema, exemplificando o procedimento acima descrito, desde a leitura da pergunta e sua etiquetação até a confecção da resposta.

3. Trabalhos futuros

Melhorias no procedimento adotado, em suas diversas etapas, podem ser realizadas. Na interpretação da pergunta, além de obter as palavras de conteúdo, como substantivos e numerais, o tipo de pergunta pode ser identificado a fim de aperfeiçoar a busca pela resposta. Assim, pode ser útil identificar se uma pergunta requer uma data ou um nome próprio, por exemplo; esse conhecimento pode ser utilizado para aumentar a precisão da resposta ao filtrar palavras que não se referem a uma data, por exemplo.

Na busca pela resposta, além do exposto acima, pode-se obter as palavras de conteúdo que estejam o mais próximo de onde ocorreu o casamento com o padrão montado com as palavras da pergunta, ao invés de montar a resposta com todos os substantivos e numerais da sentença encontrada.

Uma busca mais inteligente pode ser desenvolvida, ao invés de criar apenas uma expressão regular para a busca no conjunto de textos. Este tipo de busca pode levar a uma baixa cobertura do sistema, dado que a sentença que contenha a resposta pode não conter todos os substantivos e numerais contidos na pergunta, ou essas palavras podem aparecer em ordem diferente do exposto na pergunta.

Diversos outros aperfeiçoamentos podem ser realizados sobre a simples abordagem tratada neste texto.

4. Considerações finais

O conhecimento morfossintático obtido pela utilização do etiquetador morfossintático mostrou-se útil como um primeiro processamento do texto envolvido na tarefa de responder automaticamente a perguntas.

Salienta-se que este foi um primeiro contato com a tarefa, mostrando a potencialidade da abordagem utilizada.

Referências

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: Proceedings of the First Empirical Methods in NLP Conference. [S.l.: s.n.], 1996.

GONZALO, J. Language Resources in Cross-Language Text Retrieval: a CLEF perspective. In: Cross-Language Information Retrieval and Evaluation: Workshop of the CrossLanguage Evaluation Forum, CLEF 2000. 36—47 - Springer-Verlag – 2001.