Uma proposta de sistema de respostas automáticas a perguntas do tipo fechadas

William Augusto Rodrigues de Souza[1,2]

[1]COPPE/UFRJ – Universidade Federal do Rio de Janeiro

Caixa Postal 68511 – CEP 21945-970 - Rio de Janeiro – RJ – Brasil

[2]Divisão de Criptologia – Centro de Análises de Sistemas Navais

Praça Barão de Ladário s/n, Centro, CEP: 20091-000– Rio de Janeiro – RJ

william@cos.ufrj.br

Resumo. *Neste trabalho propomos um sistema capaz de responder automaticamente a perguntas do tipo fechadas (perguntas WH). O sistema recebe como entrada uma coleção de textos e uma ou mais perguntas e calcula a similaridade entre os textos e as perguntas, determinando os textos onde existe a maior probabilidade de ocorrência da resposta. A partir dos textos selecionados, são geradas porções menores de textos e feitos novos cálculos de similaridade. Por fim, por meio de cálculos de proximidade entre os as porções e as perguntas, a resposta é gerada.*

1. Introdução

O problema de responder perguntas automaticamente é um dos desafios atuais da Lingüística Computacional e vem sendo intensivamente estudado pelo campo de Processamento de Linguagem Natural. A tarefa consiste em produzir uma porção de texto adequada em resposta a uma pergunta feita geralmente em linguagem natural [Jurafsky e Martin 2009].

Assim, este trabalho propõe um sistema de resposta automática para perguntas fechadas (perguntas do tipo que, quem, qual, quais, onde), o qual é capaz de responder perguntas a partir de um conjunto de textos. O sistema recebe como entrada uma coleção de textos (texto plano) e um conjunto de perguntas também no formato texto e produz como saída um arquivo texto com as respostas a tais perguntas. A descoberta da resposta se dar por meio de sucessivos cálculos de similaridade em porções cada vez menores do texto. A partir de um ponto, são realizados cálculos de proximidade e é realizada uma consulta a um léxico criado a partir da coleção de textos. Por fim, um arquivo texto é gerado identificado a resposta e o texto onde a mesma foi encontrada.

2. Descrição do método proposto

Na figura 1, pode-se ver a estrutura do sistema com suas principais funcionalidades.

Modelagem. A primeira tarefa do sistema e a modelagem da coleção de textos e da pergunta em um espaço de vetores [Harman 1992]. Em determinado ponto da execução as passagens também são modeladas em um espaço de vetores.

Cálculo da similaridade. Para o cálculo da similaridade entre textos, perguntas e passagens foi utilizada a medida do ângulo do co-seno (fórmula 1) [Harman 1992].

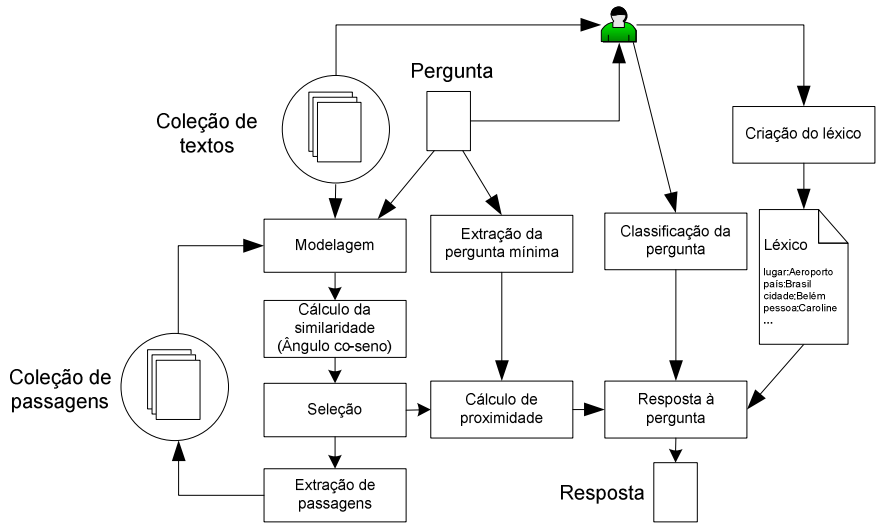

Figura 2. Estrutura do sistema proposto

Seleção. Seleciona texto ou passagens de acordo com o critério utilizado.

Extração de passagens. Extrair porções dos textos selecionados.

$$S_{Co-seno}(c_i, c_j) = \frac{\sum_{k=1}^{n}(c_{i,k} \times c_{j,k})}{\sqrt{\sum_{k=1}^{n}(c_{i,k})^2 \times \sum_{k=1}^{n}(c_{j,k})^2}} \quad (1)$$

Extração da pergunta mínima. Retira os termos irrelevantes das perguntas, deixando somente aqueles que caracterizam a necessidade de informação.

Cálculo de proximidade. Calcula a proximidade entre itens da pergunta mínima e itens nas passagens selecionadas.

Classificação da pergunta. Enquadra a pergunta em um tipo específico, considerando o pronome que a define. Por exemplo, para uma pergunta com o pronome "quem" espera-se como resposta o nome de uma pessoa.

Criação do léxico. Cria um léxico a partir da coleção de textos.

Resposta à pergunta. Responde à pergunta considerando a sua classificação, o cálculo de proximidade e o léxico criado. Um arquivo contendo a resposta é criado.

2. Descrição da ferramenta

Para a concretização do objetivo do sistema proposto foi criada a ferramenta WARS Question (figura 2). Os campos iniciais devem ser preenchidos com as informações pertinentes a localização dos diretórios e arquivos necessários.

Nas configurações do sistema pode-se indicar um arquivo de *SWords*, bem como o respectivo método de *SWord*: *Stop Word* (retira as palavras indicadas no arquivo de *SWord*), *Start Word* (só usa na coleção as palavras indicadas no arquivo de *SWord*) ou

sem *SWord*. É necessário escolher um critério para a obtenção dos textos, como, por exemplo, textos mais similares com a pergunta ou o texto com a maior similaridade com a pergunta. Existe a opção da normalização por meio de TF-IDF [Harman 1992]. É possível indicar o tamanho mínimo das palavras a considerar no processamento.

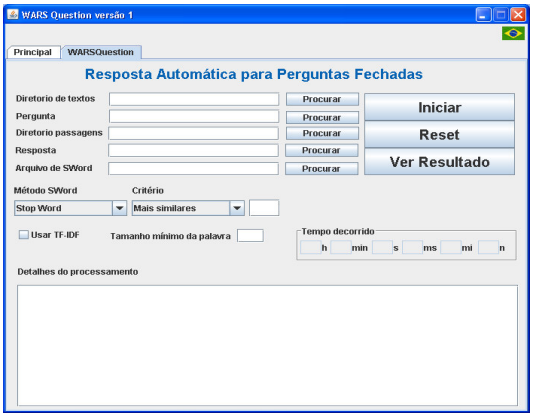

Figura 2. Interface da ferramenta WARS Question

O tempo decorrido e os detalhes do processamento são exibidos na parte inferior da interface.

3. Experimento e resultado[1]

O objetivo do experimento é aplicar o sistema proposto a um conjunto de 20 textos para responder a 30 perguntas proposta em [IOlinCom 2009]. O sistema obteve sucesso respondendo corretamente as 30 perguntas propostas, com duas incorreções relacionadas à ordem alfabética dos termos em duas respostas.

4. Conclusões

O trabalho propõe um sistema e demonstra sua eficácia para responder a um conjunto de perguntas propostas a partir de uma coleção de textos. O experimento descreveu que o sistema respondeu corretamente as perguntas gerando apenas duas incorreções relacionadas à ordem alfabética dos termos em duas respostas.

Referências Bibliográficas

IOlinCom, I Olimpíada Brasileira de Lingüística Computacional (2009). Disponível em: http://www.nilc.icmc.usp.br/~arianidf/olincom/trilha1.html. [acessado13 abr. 2009].

Harman, D. (1992), "Ranking algorithms". In Information retrieval: data structures and algorithms, Edited by William Frakes and Ricardo Yates, Prentice Hall, p. 363–392.

Jurafsky, D. and Martin, J. H. (2009), Speech and language processing: An introduction to natural language processing, computacional linguistics, and speech recognition. 2th ed. Pearson.

Yates, R.B. and Neto, B. R. (1999), Modern information retrieval. Addison Wesley.

[1] Mais detalhes sobre esse trabalho podem ser obtidos em http://www.portalcomputacao.com.br