SQAS: Um Sistema Automático de *Question-Answering* para Textos Jornalísticos

Danilo Machado Junior[1], Juliano Henrique Foleiss[1], Vinícius Mourão Alves de Souza[1]

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
Maringá – PR – Brazil

{danilo.junior,juliano.foleiss,vsouza}@din.uem.br

Abstract. *This paper describes the building of SQAS (Shallow Question-Answering System), an automatic Question-Answering (QA) system for hard-questions for Portuguese journalistic papers. To accomplish this goal, the system is broken into two modules: the first one search the texts to find the best paragraph candidate to hold the answer and the second one scans the candidates for the answer using regular expressions defined for each type of hard-question. The system proved satisfactory on finding the right paragraphs and on scanning for certain types of answers.*

Resumo. *Este artigo descreve a construção do SQAS* (Shallow Question-Answering System*), um sistema automático de respostas para perguntas do tipo* "hard questions" *aplicado a textos jornalísticos em português. Com a finalidade de cumprir essa tarefa, o sistema utiliza dois módulos distintos: o primeiro busca textos e parágrafos candidatos e o segundo filtra as possíveis respostas de acordo com padrões definidos para cada tipo de pergunta. O sistema se mostrou preciso na definição dos parágrafos que contêm a resposta e na filtragem de alguns tipos de pergunta.*

1. Introdução

A construção de sistemas de perguntas e respostas (*Question-Answering*) é uma das tarefas da extração de informação (*Information Retrieving*), na área da Lingüística Computacional. Tais sistemas devem responder a perguntas formuladas em linguagem natural, formulando as respostas a partir de uma coleção de textos (Harabagiu & Maldovan, 2005; Manning *et al.*, 2008).

O objetivo deste artigo é descrever o SQAS *(Shallow Question-Answering System*), um sistema de perguntas e respostas aplicado a textos jornalísticos em português. Tal sistema foi construído para obter respostas para trinta perguntas de tipo fechado, propostas como uma das tarefas de avaliação da I OLinCom[1]. As respostas são buscadas em um corpus composto por vinte textos jornalísticos disponibilizado como parte da tarefa de avaliação.

[1] A I Olimpíada Brasileira de Linguística Computacional (I OLinCom) é uma competição científica vinculada ao Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2009). Disponível em http://www.nilc.icmc.usp.br/~arianidf/olincom/

A Seção 2 descreve a construção e funcionamento do sistema SQAS. A Seção 3 apresenta e discute os resultados obtidos pelo sistema e, finalmente, a Seção 4 apresenta as conclusões.

2. O sistema SQAS

O sistema SQAS foi construído utilizando-se a linguagem de programação C, no sistema operacional Linux, e pode ser dividido em dois módulos distintos, descritos a seguir.

O primeiro módulo define quais são o texto e o parágrafo que possuem maior probabilidade de conter a resposta por meio da contagem da freqüência das palavras presentes na pergunta nesses textos/parágrafos. Antes da contagem das freqüências, são retiradas das perguntas as palavras de pouca relevância (*stopwords*).

O segundo módulo refina a possível resposta, retirando dos parágrafos selecionados as frases que não contenham nenhuma palavra da pergunta. Para as perguntas do tipo "quem", "quanto" e "onde", padrões de respostas exatas são procurados por meio de expressões regulares. Para os outros tipos de questões, a resposta apresentada é o parágrafo refinado produzido pelo segundo módulo.

```
Início:
Carregar todas as questões, retirando as stopwords
Carregar todos os textos

Fase 1: Determinar em qual texto e qual parágrafo está a resposta
  Contar a frequência das palavras da pergunta nos textos (I)
  Verificar, para cada questão, qual é o parágrafo mais relevante entre todos os textos,
  ou seja, qual é o parágrafo dentre todos os textos que contém a maior quantidade de
  palavras da pergunta. Em caso de empate, o programa escolhe o primeiro. (II)
  Determinar que a resposta está
     No texto com a maior frequência (medida I)
     No parágrafo mais relevante (medida II)

Fase 2: Determinar a resposta
  Para cada pergunta
     Selecionar o parágrafo com a resposta
        Separar o parágrafo em sentenças
        Para cada sentença
           Verificar se ela contém alguma palavra da pergunta. Se sim, inclua a sentença
           no parágrafo de sentenças relevantes
     Enquanto a chamada ao analisador léxico correspondente ao tipo da pergunta (usando
     como entrada o parágrafo com sentenças relevantes recém criado) retornar alguma
     resposta candidata
        Incluir a resposta candidata na lista de respostas candidatas
     Escolher o tipo da pergunta (QUAL, QUE, QUANTO, ONDE, QUEM)
        Se a pergunta for QUAL ou QUE
           Não procurar a resposta exata. Apenas retornar o parágrafo todo
        Caso contrário (pergunta QUANTO, ONDE, QUEM)
           Para cada resposta candidata na lista de respostas candidatas
              Se alguma palavra da resposta candidata for igual a  qualquer  palavra  da
              pergunta, retirar essa resposta da lista de respostas candidatas
           Para cada resposta candidata na lista de respostas candidatas
              Calcular a posição que a resposta candidata se encontra no parágrafo de
              sentenças relevantes (IV)
           Para cada palavra da pergunta
              Calcular a posição média dela na resposta (posição acumulada das
              ocorrências da palavra/número de ocorrências da palavra) (V)
           Para cada resposta candidata na lista de respostas candidatas
              Calcular a distância entre a posição média de cada palavra da pergunta na
              resposta (medida V) e a posição da resposta candidata (medida IV) (VI)
           Escolher a resposta com a menor distância (medida VI) como a resposta
```

Algoritmo 1. SQAS

O Algoritmo 1 apresenta os mecanismos implementados pelo sistema para o primeiro módulo (representado no algoritmo como Fase 1) e para o segundo módulo (representado no algoritmo I como Fase 2).

3. Resultados e Discussão

Na I OLinCom, a avaliação dos sistemas foi feita por meio da atribuição de notas de um a quatro de acordo com a corretude das respostas fornecidas e da indicação dos textos corretos apontados pelo sistema. O SQAS obteve sessenta e oito pontos dos cento e vinte possíveis, sendo que o sistema tem um bom desempenho em encontrar o texto e o parágrafo em que se encontram a resposta correta, porém o desempenho do módulo de refinamento da resposta, que busca pela *string* de resposta específica, ainda está aquém do desejado.

A pontuação obtida pelo SQAS poderia ser melhorada com a inclusão de um desempate mais criterioso de parágrafos com a mesma relevância na seleção das respostas. Além disso, uma busca de padrões e a construção de expressões regulares para os tipos de questões que não sejam "quem", "quanto" e "onde", atualmente suportados, poderiam melhorar sensivelmente o desempenho do sistema e, conseqüentemente, sua pontuação.

4. Conclusão

Este artigo apresentou o sistema SQAS, um sistema de perguntas e respostas aplicado a textos jornalísticos em português. Tal sistema foi construído com parte das avaliações da I OLinCom. Para as perguntas fornecidas como parte da tarefa de avaliação, o sistema é capaz de encontrar o texto e o parágrafo em que se encontram a *string* de resposta com um bom desempenho, entretanto, melhorias no módulo de busca da *string* de resposta dentro do parágrafo são necessárias.

Referências

Harabagiu, S.; Maldovan, D. (2005). Question Answering. In Ruslan Mitkov (Ed.) *The Oxford Handbook of Computational Linguistics*. Oxford University Press, pp. 560-582.

Manning, C.D.; Raghavan, P.; Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.