Processamento Eficiente de Consultas em Linguagem Natural em Ambientes Educacionais

Antônio Luiz Mattos de Souza Cardoso[1,2], Crediné Silva de Menezes[2]

[1]Centro de Desenvolvimento de Sistemas de Vitória (CDSV) – Xerox do Brasil
Av. Fernando Ferrari, 1000 – 29.060-410 – Vitória – ES – Brasil

[2]Programa de Pós graduação de Informática – Universidade Federal do Espírito Santo (UFES)
Av. Fernando Ferrari, s/n – 29.060-900 – Vitória – ES – Brasil

antonio.cardoso@bra.xerox.com, credine@inf.ufes.br

Abstract*. Automated Query and Answering systems need tools which help their users to formulate questions and have their answers processed fast and accurately. This paper describes a tool, based on natural language processing, that helps a 'human answerer' to answer different questions. This tool identifies and clusters various similar questions formulated in natural language, so that the 'human answerer' may answer all clustered questions at once. This tool has been developed applying different NLP resources and techniques, such as: Stemming, Query expansion, Thesaurus, Grammar rules of the Portuguese language, Relevance feedback, and Proper names identification*

Resumo. *Sistemas automatizados de Perguntas e respostas necessitam de ferramentas que auxiliem na elaboração de consultas e no processamento de suas respostas de modo preciso e rápido. Este artigo descreve uma ferramenta que auxilia o 'respondedor humano' na sua tarefa de responder a diferentes consultas. Esta ferramenta identifica e agrupa diferentes consultas similares entre si elaboradas em linguagem natural, de modo que o 'respondedor humano' possa responder simultaneamente a todas as consultas agrupadas. Ela foi desenvolvida utilizando diferentes técnicas de NLP, tais como: Stemming, Query expansion, Dicionário de sinônimos, Regras gramaticais da língua portuguesa, Relevance feedback e Identificação de nomes próprios.*

1. Introdução

Indivíduos que buscam a mesma informação normalmente utilizam formas lingüísticas diferentes para formular as suas consultas. As diferenças entre consultas podem estar nos termos utilizados, na quantidade dos termos, nas variações sintáticas ou semânticas dos termos. Apesar de serem diferentes na sua formulação, as consultas são similares pois, buscam a mesma informação.

Este artigo descreve uma ferramenta que identifica e agrupa diferentes consultas, similares entre elas, formuladas em linguagem natural. O agrupamento das consultas similares permite que elas sejam respondidas simultaneamente por uma única resposta, tornando o sistema mais eficiente.

2. Ambientes Educacionais

Um sistema de Perguntas e Respostas, que armazena as consultas dos usuários não respondidas para serem processadas posteriormente, pode apresentar um acúmulo delas. O acúmulo pode provocar um gargalo no sistema reduzindo a sua capacidade de processar as consultas, tornando-o ineficiente. Assim, um usuário, que formula uma nova consulta neste contexto, pode esperar mais tempo pela sua resposta. Como solução, aumentar a capacidade de processamento do sistema pode ser inviável devido a custos financeiros ou devido a solução já ter alcançado seus limites de expansão de recursos, entre outras razões.

2.1. O AmCorA

O AmCorA é um projeto de ambiente inteligente e cooperativo de aprendizagem por computador baseada na teoria construtivista em desenvolvimento na Universidade Federal do Espírito Santo. O AmCorA possui um repositório de perguntas e respostas de modo que responde de forma automática a consultas formuladas por seus usuários. Assim, a cada nova consulta formulada, ele verifica o repositório de perguntas e respostas e devolve ao usuário uma resposta que julga ser adequada àquela consulta. Caso o sistema não encontre no repositório a resposta adequada para uma determinada consulta, ele a direciona para um colaborador humano, que o AmCorA determina ser o mais capacitado ao tipo da pergunta, para respondê-la.

3. Proposta de solução

Para solucionar o problema de acúmulo de consultas não respondidas, foi proposta uma solução que agrupe aquelas que sejam similares através da morfologia das consultas.

3.1. Agrupamento de consultas por morfologia

Este recurso aplica técnicas de processamento de linguagem natural para identificar e agrupar consultas similares a fim de o usuário responder uma única vez a todas as consultas que foram consideradas como similares e agrupadas pela solução. Entre as funcionalidades há: Identificação e classificação do tipo da consulta; Identificação de nomes próprios e frases padrão; Análise léxica; Eliminação de stopwords; Aplicação de stemming; Execução de query expansion e Obtenção de relevance feedback do usuário.

A *identificação e classificação do tipo da consulta* é uma importante funcionalidade deste primeiro recurso pois é iniciado o processo de entendimento da necessidade de informação do usuário. O pronome ou advérbio interrogativos, existente na consulta, revela o tipo de resposta pretendido pelo usuário [Wen and Nie 2002].

A *identificação de nomes próprios e frases padrão* segue a premissa de que o reconhecimento deles melhora a performance da recuperação da informação [Thompson and Dozier 1999]. São identificados nomes de pessoas, organizações e localidades. Para a identificação de nomes próprios, foram definidas regras heurísticas combinadas com tabelas de nomes primários. As regras foram implementadas com código recursivo. Para a identificação de frases padrão, foi criada uma tabela contendo frases e expressões que possuem significado único num determinado contexto. Primeiramente, é identificado todos os nomes de organizações, depois os nomes de pessoas e por fim os nomes de localidades. Frases padrão são cadastradas pelos usuários permitindo que elas não sejam processadas por nenhuma função, ou seja, elas não sofrem redução a raiz gramatical, remoção de acentuação, comparação com sinônimos ou qualquer outra operação existente na solução.

A *análise léxica do texto* identifica as palavras contidas na consulta. Quatro situações especiais são tratadas neste momento: dígitos, hífens, palavras em letras maiúsculas/minúsculas e sinais de pontuação e acentuação. Os dígitos não são considerados como relevantes em sistemas de recuperação da informação [Yates and Neto 1999]. Pela mesma razão, sinais de acentuação e hífens são também removidos das palavras. Todas as palavras são convertidas para maiúsculas. E, ao término da execução desta funcionalidade, os sinais de pontuação são também removidos das consultas.

A *remoção de stopwords* remove as palavras que não carregam significado contidas nas consultas [Yates and Neto 1999]. Assim, uma tabela dinâmica com 320 palavras foi criada contendo artigos, pronomes, preposições, interjeições e conjunções.

A *aplicação de stemming* reduz os termos das consultas a sua raiz gramatical pela remoção de prefixos e sufixos. O algoritmo de stemming é baseado em 220 regras gramaticais da língua portuguesa com 8 etapas [Orengo and Huyck 2001]. Uma etapa adicional, Redução de Advérbio, foi inserida na seqüência original a fim de que palavras terminadas em “mente”, podem ser reduzidas a sua raiz gramatical.

A *execução de query expansion* incorpora novos termos sinônimos aos termos originais da consulta. Isto permite que as consultas, que possuem termos diferentes na sua composição, possam ser consideradas similares caso elas possuam termos sinônimos. Para que query expansion fosse aplicado na solução, foi criada manualmente uma tabela de palavras e seus sinônimos da língua portuguesa utilizando um dicionário de sinônimos com 30.000 verbetes [Fernandes 1999].

Finalmente, *relevance feedback do usuário* é obtida pela apresentação do resultado do agrupamento das consultas similares gerado pela solução. O usuário decide se a qualidade do agrupamento está adequada. Caso não esteja, o usuário pode tentar novamente reconfigurando os parâmetros Grau de similaridade, Tipo da consulta e Query expansion para obter um outro conjunto. O parâmetro Grau de similaridade permite aos usuários informarem qual é a percentagem de termos similares, originais ou

sinônimos, no conjunto de consultas agrupadas que a solução deve apresentar. O Grau de similaridade varia entre 50 a 100%.

4. Considerações Finais

A combinação de diferentes técnicas de processamento de linguagem natural possibilitou o desenvolvimento de uma solução que agrupa diferentes consultas, formuladas em linguagem natural e similares entre si, permitindo que elas possam ser respondidas simultaneamente. O agrupamento das consultas é totalmente automatizado.

4.1. Testes e Resultados

Testes de sistema foram realizados na solução. Participaram dos testes, engenheiros de software do CDSV que formularam 200 consultas sobre esporte, religião e política. Além de formularem as consultas, eles identificaram quais os diferentes agrupamentos de consultas similares existiam entre as consultas a fim de comparar com o agrupamento das consultas similares gerado pela solução com o agrupamento humano.

Apesar de os testes não terem sido exaustivos, a primeira conclusão obtida pelos usuários foi a facilidade de uso. Os usuários, sem conhecimentos sobre técnicas e ferramentas de NLP, agruparam consultas similares rapidamente. Além da facilidade e rapidez, os seguintes testes e resultados foram obtidos:

- **Grau de Similaridade igual a 100% e Query Expansion ativo**: todas as consultas similares dos diferentes agrupamentos foram agrupadas corretamente;
- **Grau de Similaridade menor que 100% e Query Expansion ativo**: todas as consultas dos diferentes agrupamentos foram agrupadas corretamente. Algumas consultas não pertinentes aos agrupamentos foram agrupadas; e
- **Grau de Similaridade igual a 100% e Query Expansion desativado**: Algumas consultas similares dos diferentes agrupamentos não foram agrupadas.

Através dos resultados, foi verificado que o comportamento da solução é exatamente o esperado. Os agrupamentos de consultas não esperados são gerados pela solução a fim de flexibilizá-la para que o usuário decida o que ele deseja agrupar.

5. Referências

Fernandes, F. “Dicionário de Sinônimos e Antônimos da Língua Portuguesa”, 38. ed., São Paulo, Editora Globo, 1999.

Orengo, V. and Huyck, C. (2001) “A Stemming Algorithm for the Portuguese language”, In: Proceedings of SPIRE 2001. IEEE Computer Society.

Thompson, P. and Dozier, C. (1999) “Name Recognition and Retrieval Performance”. Natural Language Information Retrieval. Netherlands: Kluwer Academic Publishers, p. 261-272.

Wen, J. and Nie, J. (2002) “Query Clustering Using User Logs”, ACM Transactions on Information Systems, v. 20, n. 1, p. 59-81.

Yates, R. and Neto, B. “Modern Information Retrieval”, New York, Addison-Wesley, 1999.