RBE - Caderno de Engenharia Biomédica, v. 13, n. 3, p. 57-68, julho 1997

Edição Especial: Engenharia Biomédica na América Latina

PROCESSAMENTO DA LINGUAGEM NATURAL APLICADO À EXTRAÇÃO AUTOMÁTICA DE INFORMAÇÕES SOBRE AVALIAÇÃO DE TECNOLOGIAS EM SAÚDE

S. M. Freire[1] e R.B.Panerai[2]

RESUMO -- Este trabalho tem como objetivo o de avaliar a possibilidade de se desenvolver um sistema computacional para extrair automaticamente informações sobre tecnologias em saúde a partir da leitura de resumos publicados na literatura médica. Uma amostra de resumos de artigos relacionados à indução do parto e/ou amadurecimento cervical foi recuperada do MEDLINE e dividida em dois grupos. Um grupo foi utilizado para análise e desenvolvimento do sistema. O outro grupo foi usado para testar o sistema quanto à capacidade de extrair a partir do título do resumo as tecnologias avaliadas em cada estudo. O programa produziu informações corretas em 71,1 % dos resumos. Este estudo mostra o potencial da utilização de técnicas de interpretação da linguagem natural para extrair dos resumos médicos informações sobre tecnologias em saúde e implementar uma base de dados voltada para Avaliação de Tecnologias em Saúde (ATS).

Palavras-chave: Processamento da Linguagem Natural, Extração de Informações, Literatura Médica.

INTRODUÇÃO

A literatura médica é uma das melhores fontes de informação sobre os efeitos do uso de tecnologias em saúde. O crescente número de revisões sistemáticas da literatura (meta-análises) reflete o reconhecimento da sua importância para melhorar o conhecimento sobre os efeitos dos cuidados com a saúde. O *Cochrane Collaboration* - uma rede internacional de indivíduos e instituições - evoluiu para produzir revisões sistemáticas, periodicamente atualizadas, de ensaios clínicos randomizados (ECR) (Chalmers e Haynes, 1994). A magnitude desta tarefa não deve ser subestimada: parece provável que, mesmo com a atenção focalizada somente nos ensaios clínicos randomizados, cerca de 1 milhão de estudos realizados durante a segunda metade do século XX

[1] Programa de Engenharia Biomédica - COPPE/UFRJ e FUNREI - Fundação de Ensino Superior de São João del Rei - MG
Posição Atual: Professor Adjunto - Informática Médica/LAMPADA/FCM/UERJ, Rua Prof. Manuel de Abreu 48/2º andar, CEP 20560-001, Vila Isabel, Rio de Janeiro,
e-mail: sfreire@saude.lampada.uerj.br
[2] Programa de Engenharia Biomédica - COPPE/UFRJ
Posição Atual: Senior Lecturer, Department of Medical Physics, Leicester University, LE1 5WW, Leicester, England, e-mail: rp9@leicester.ac.uk

// **Trabalho recebido em 28/02/96 e aceito em 18/10/96** //

devem ser considerados. É provável que será preciso algumas décadas para um estado estável ser atingido, no qual resultados de novos estudos primários vão sendo incorporados eficientemente nas revisões sistemáticas de pesquisas (Chalmers e Haynes, 1994).

Os ensaios clínicos randomizados não são, porém, a única fonte de informações sobre Avaliação de Tecnologias em Saúde (ATS). Outros delineamentos de pesquisa, tais como ensaios clínicos não randomizados, estudos de coortes, estudos de caso-controle etc, podem fornecer informações úteis. Se levarmos em consideração tais estudos, as observações de Chalmers e Haynes são ainda mais pertinentes.

Um instrumento que tornaria mais fácil a tarefa de revisão da literatura seria uma base de dados estruturada de tal forma que estudos que comparam as mesmas tecnologias com a mesma indicação de uso e com delineamento semelhante estejam agrupados e sejam acessados por um sistema dirigido por menus. Um exemplo é mostrado na figura 1. No primeiro menu, possíveis problemas de saúde são apresentados. A escolha de um item (indução do parto no caso) leva a um outro menu que apresenta uma lista de possíveis comparações entre as tecnologias que são empregadas no problema anterior. Escolhendo uma destas comparações (PGE2 X Oxitocina), o usuário é, então, guiado para um terceiro menu que mostra possíveis delineamentos de estudos comparativos. A seleção de um tipo de delineamento resulta então em uma lista de estudos que comparam as tecnologias escolhidas de acordo com o delineamento e problema de saúde selecionados.

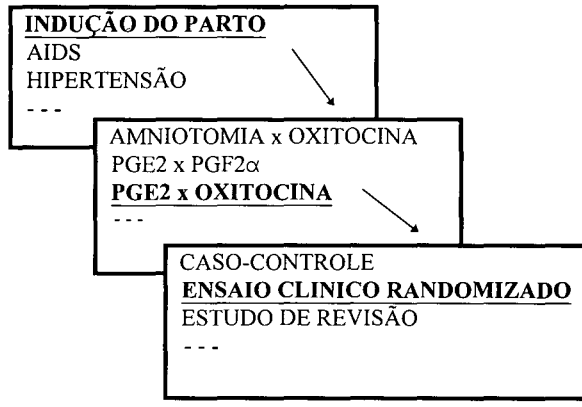

Figura 1. Um sistema dirigido por menus para uma base de dados voltada para ATS.

Com tal base de dados orientada para tecnologias, poderemos ter melhores medidas de precisão e recuperação de estudos de avaliação tecnológica. O número de revisões da literatura deveria estar limitado somente pelo número de pessoas disponíveis para conduzí-las, já que a etapa de busca de estudos na literatura seria bastante facilitada com bases voltadas para tecnologias. Além disso, esta base poderia proporcionar um mapeamento do campo de ATS, indicando áreas que são objeto de pouca ou nenhuma avaliação e que, por isso, requerem maior atenção por parte dos

planejadores dos sistemas de saúde. Em países como o Brasil, onde a área de avaliação tecnológica em saúde ainda é incipiente, uma base voltada para a ATS contribuiria de maneira eficiente para o acesso a informações que auxiliem na decisão clínica, gerenciamento hospitalar e política de alocação de recursos. Esse instrumento seria extremamente útil para apoiar as atividades de coordenação de um órgão de avaliação tecnológica que eventualmente venha a ser criado no país, seguindo a tendência dos países desenvolvidos.

Por outro lado, a construção de tal base de dados apresenta alguns dos problemas que temos com os trabalhos de revisão da literatura se a busca de estudos de avaliação tiver que ser realizada manualmente, como vem sendo o caso do *Oxford Database of Perinatal Trials* (ODPT) (Chalmers, 1988). Um grande avanço nesta área seria um sistema que varresse automaticamente as bases tradicionais e automaticamente implementasse uma base voltada para tecnologias com as características mencionadas nos parágrafos anteriores.

Técnicas de Processamento da Linguagem Natural atingiram um estágio onde aplicações em diversas áreas estão sendo testadas. Em Medicina, o interesse maior se concentra em interfaces com banco de dados, indexação de trabalhos científicos (Evans *et alii*, 1991) e análises de prontuários médicos (Rassinoux *et alii*, 1994; Ledoray *et alii*, 1992; Sager *et alii*, 1987). Este artigo apresenta uma outra aplicação: um enfoque para ler artigos e extrair informações dos mesmos usando técnicas de processamento da linguagem natural.

MATERIAIS E MÉTODOS

Para investigar o potencial dos resumos para fornecer informações relativas à avaliação tecnológica em saúde, uma amostra de 187 resumos de artigos relacionados com a indução do parto e/ou amadurecimento cervical foi extraída do MEDLINE, no período entre 1986 e 1990. Os descritores utilizados foram parto (*labor, labour*) e indução (*induced*). Este tópico foi escolhido por duas razões básicas. A primeira é que este é um problema de saúde sobre o qual muitos ensaios clínicos têm sido realizados, conforme pode ser comprovado através de uma inspeção no *Oxford Database of Perinatal Trials* (ODPT). Uma segunda razão é que o problema da indução do parto e/ou amadurecimento cervical está restrito a um breve período da gestação, requer poucas tecnologias e o conhecimento que deve ser representado, no que concerne às informações sobre ATS, é pequeno quando comparado a outros problemas de saúde.

Para implementar uma base de dados voltada para tecnologias, as seguintes informações devem ser coletadas: título, autores, periódicos (volume, número e páginas), tecnologias avaliadas, tecnologia padrão, indicações e condições de uso das tecnologias, delineamento do estudo e grupos de pacientes avaliados. O título, autores e informações relativas ao periódico são facilmente extraídos das bases tradicionais.

Um sistema computacional foi implementado para extrair automaticamente as tecnologias avaliadas a partir do título dos artigos da literatura médica. O sistema, chamado **SAARME** (Sistema para **A**nálise **A**utomática de **R**esumos **ME**dicos) consiste dos seguintes componentes: um dicionário ou léxico e um analisador sintático-semântico.

Para o desenvolvimento do SAARME, a amostra de 187 resumos foi aleatoriamente dividida em dois grupos: um grupo (grupo A), consistindo de 84 resumos, foi usado para construir o léxico e desenvolver o sistema; o segundo grupo (grupo B), com 83 resumos, foi o grupo de teste. Um dos autores realizou uma leitura dos resumos do grupo A para verificar a extensão com que as informações listadas acima podem ser extraídas do texto dos resumos.

Léxico

O léxico é um arquivo contendo as palavras que o sistema conhece. Para cada vocábulo, suas classes sintáticas e semânticas são fornecidas. Um conjunto de classes sintáticas e semânticas foram criadas e atribuídas às palavras do dicionário. As classes sintáticas ajudam a construir as estruturas sintáticas das sentenças dos títulos; elas estão relacionadas com as classes sintáticas da gramática inglesa. As classes semânticas são empregadas para mapear as informações extraídas dos títulos em um mecanismo de representação, de modo que possam mais tarde serem recuperadas, se for necessário. As classes semânticas foram criadas de acordo com o domínio enfocado pelo sistema: Medicina e Avaliação Tecnológica em Saúde.

Como exemplo, a palavra *Oxytocin* tem a seguinte entrada no dicionário:
OXYTOCIN - N - HORM

O sistema interpreta esta palavra como tendo a classe sintática substantivo (N) e a classe semântica Hormônio (HORM). Todas as palavras do grupo de estudo foram classificadas e inseridas no dicionário.

O Analisador Sintático-Semântico

A análise sintática é baseada em um analisador sintático determinístico (Marcus, 1980). À medida que cada palavra do título é lida, ela é inserida em alguma estrutura sintática. Ao final, é obtida uma estrutura que representa o conteúdo sintático do título e que servirá como entrada para o analisador semântico. Este analisador verifica o significado semântico de cada palavra no núcleo de sintagmas nominais. A cada classe semântica que possa gerar informações relevantes sobre tecnologias está associada uma rotina que analisa os modificadores à direita e à esquerda do núcleo do sintagma nominal e constrói objetos semânticos relacionados a estes modificadores. O objeto associado à estrutura principal é inserido em uma lista encadeada, de agora em diante chamada LISTA. Após todas as palavras serem analisadas, os componentes de LISTA são analisados para verificar se eles se encaixam em algum padrão através do qual as informações transmitidas são reconhecidas pelo sistema.

O sistema foi implementado em C++. Para cada estrutura sintática ou semântica, uma classe foi definida e, quando tais estruturas são geradas, um objeto da classe correspondente é criado e as palavras são ligadas aos campos apropriados do objeto. O mesmo se aplica à representação final das tecnologias avaliadas. Para mostrar como o sistema opera, vejamos os principais passos da análise da título: *Preinduction cervical priming in high risk pregnancy. Experience with a new sustained release PGE2 vaginal film.*

A análise sintática da primeira parte do título gera um sintagma nominal e um sintagma preposicional que são mapeados nos seguintes objetos:

NSTG (NP1)
NUCLEO (N1)
N - *priming*
LM (LM1)
APOS - *cervical*
NPOS (NPOS1)
N - *induction*
LCDN - *pre*
PP (PP1)
PREP - *in*
NSTG (NP2)
NUCLEO (N2)
N - *pregnancy*
APOS - *high*
NPOS (NPOS2)
N - *risk*

Na representação acima, cada nível de indentação contém os componentes de um objeto que está no nível imediatamente superior. Os identificadores associados às palavras do texto são campos que compõem o objeto ao qual eles pertencem. A cada objeto criado está associado um identificador, representado entre parênteses. O significado de cada classe ou campo é especificado abaixo:

NSTG - sintagma nominal
NUCLEO - núcleo do sintagma nominal
PP - sintagma preposicional
APOS - adjetivos
N - substantivos
NPOS - substantivos que modificam o núcleo de um sintagma nominal
PREP - preposição
LM - modificadores da palavra que compõe o núcleo
LCDN - modificadores do substantivo que compõe NPOS

Assim o sintagma nominal NP1 é composto de um NUCLEO, N1, modificado à esquerda por um objeto do tipo NPOS (NPOS1) e à direita por um sintagma preposicional, e assim por diante.

A análise semântica da estrutura sintática gerada se inicia pelo reconhecimento da classe semântica da palavra *priming*. Esta classe, procedimento, leva à criação de um objeto do tipo PROC (PROC1), com o campo nome preenchido com a palavra *priming*. Este objeto é ligado a um outro objeto do tipo TECH (TECH1) que significa tecnologia. A análise dos modificadores à esquerda leva à criação de outros objetos ligados ao procedimento PROC1. A representação final de TECH1 é mostrada a seguir. Os objetos BODY1 (PARTECORPO) e ORG1 (ORG - órgão) são auto-explicativos. O objeto TEMP1 (TEMPOEVENTO) indica que este procedimento é aplicado antes (indicado por trel = *pre*) do procedimento PROC2:

TECH (TECH1)
PROC (PROC1)
nome - *priming*
PARTECORPO (BODY1)
ORG (ORG1)
nome - *cervix*
TEMPOEVENTO (TEMP1)
trel - pre
PROC (PROC2)
nome - *induction*

Como estamos procurando informações sobre as tecnologias avaliadas em cada estudo, o conteúdo semântico do sintagma preposicional que modifica à direita o sintagma nominal anterior não leva à construção de nenhuma estrutura. O objeto TECH1 é inserido em LISTA, seguido por um ponto ('.').

A segunda parte do título é então analisada e TECH2 é criado com a seguinte estrutura:

TECH (TECH2)
HORM (HORM1)
nome - *prostaglandin*
categoria - *E2*
USO (USO1)
PARTECORPO (BODY2)
ORG (ORG2)
nome - *vaginal*
veiculo - *film*

TECH2 é então inserido em LISTA, cujos componentes são agora TECH1, '.' e TECH2. Esta sequência se encaixa em um padrão que é identificado por SAARME. Uma vez que TECH2 é um hormônio e TECH1 é um procedimento, o sistema estabelece que TECH2 é a tecnologia avaliada e que TECH1 é o objetivo de uso da tecnologia expressa em TECH2.

Os outros títulos são analisados de modo semelhante. Alguns dos padrões identificados por SAARME são mostrados abaixo (outros padrões são apresentados em Freire (1994)):

1) TECH:	tecnologia avaliada;
2) TECH1 for TECH2:	TECH1 é avaliada com o objetivo expresso por TECH2;
3) TECH1 by TECH2:	TECH2 é avaliada com o objetivo expresso por TECH1;
4) TECH1 + TECH2 for TECH3:	TECH1 é usada com TECH2 para o objetivo TECH3.

RESULTADOS

As informações que podem ser extraídas de cada resumo do grupo de estudo estão listadas com detalhes em Freire (1994). Um resumo destes resultados é apresentado a seguir.

As informações relativas às tecnologias (tecnologia avaliada, tecnologia padrão, grupos de pacientes e condições de uso das tecnologias) são fornecidas na maioria dos textos, embora algumas restrições mereçam atenção. Em alguns resumos, as informações não são oferecidas de maneira explícita, sendo que as mesmas devem ser inferidas a partir da leitura do texto, o que pode gerar dúvidas e dar margem à subjetividade. Em outros casos, as informações estão misturadas como, por exemplo, quando a tecnologia padrão é apresentada somente na parte de resultados.

A informação sobre a condição de uso das tecnologias não é apresentada de forma consistente nos resumos. Neste trabalho, o critério utilizado para obter estas informações (característica dos pacientes avaliados) implica que a qualidade desta informação depende do grau de especificação do estado dos pacientes dado pelos autores do resumo. Como pode ser observado, a qualidade desta especificação varia desde informações vagas ("medical reasons") até um grau elevado de detalhamento (*high risk primiparae with very poor cervical scores (less than 3)*).

Quanto à informação sobre o delineamento do estudo, ocorre o mesmo problema que aquele observado em relação à indicação de uso tecnológico. Em geral, os ensaios clínicos randomizados, pelo seu status junto à comunidade científica, são claramente especificados nos resumos, e até mesmo nos títulos. Às vezes, estudos de casos e de revisão da literatura apresentam esta informação nos resumos; em outros casos esta informação pode ou não ser deduzida da leitura do texto. Estudos transversais, controlados ou não, e estudos descritivos não apresentam informação sobre o delineamento.

Se nos limitarmos à leitura do título, a única informação que pode ser extraída com certa confiança é a tecnologia avaliada (77,6% dos casos). Mas mesmo assim, uma leitura do texto do resumo é necessária para conferir a veracidade da informação.

O programa SAARME foi rodado com o segundo grupo, usando dois dicionários distintos. O primeiro dicionário era o básico, consistindo das palavras dos resumos do grupo A. O segundo dicionário era uma extensão do primeiro, com as novas palavras dos títulos e resumos do grupo B adicionadas ao dicionário básico. Os resultados são apresentados na tabela 1. Os resultados foram comparados com os obtidos da leitura somente dos títulos dos resumos por um dos autores.

Tabela 1: Resultados do teste com SAARME.

	Identificação Correta	Identificação Errada	Percentual (%)	Total
Dicionário Básico	32	51	38,6	83
Dicionário Estendido	59	24	71,1	83

Usando o dicionário básico, a percentagem de identificações corretas foi somente de 38,6% (32/83). Com o dicionário estendido, o sistema identificou corretamente a tecnologia em 71,1% (59/83). Neste último caso, a não-identificação se deve principalmente à análise sintática errada do texto (15 títulos), e também à análise semântica (9 títulos).

DISCUSSÃO

Os resumos dos artigos científicos são de grande importância para verificar de forma rápida o conteúdo de um trabalho. Este estudo se propôs a utilizar os resumos também para a implementação automática de uma base de dados voltada para ATS. Inicialmente o sistema implementado, SAARME, analisa os títulos dos resumos. Nesta seção, consideraremos os seguintes aspectos: discussão do SAARME, análise dos corpos dos resumos, outros enfoques, padronização de informações.

Na área de ATS, o trabalho realizado pelo Cochrane Collaboration é de grande importância. As bases desenvolvidas por esta organização permitem encontrar informações sobre eficácia de tecnologias obtidas através de ensaios clínicos randomizados. Entretanto, outros parâmetros de avaliação tecnológica como custo, efeitos colaterais e efetividade raramente são analisados em ECRs. Assim, para buscar estas informações, deve-se recorrer às bases tradicionais; o mesmo se aplica às tecnologias avaliadas por outros métodos além do ECR. Desta forma, um sistema automático que organize a literatura relativa à ATS viria a auxiliar o esforço do Cochrane Collaboration no que relaciona à busca de ensaios clínicos randomizados e a complementá-lo quando se trata de outros métodos de avaliação tecnológica.

O SAARME deve ser visto como um protótipo. Ele não se propõe, no estágio presente, a realizar buscas bibliográficas e nem deve ser comparado com outros sistemas de recuperação. O que esse trabalho pretendeu foi mostrar as vantagens de um sistema automático para a construção de uma base de dados voltada para ATS e a implementação do SAARME como um passo nessa direção. Pesquisas relacionadas à interpretação da linguagem natural, sistemas de recuperação bibliográfica e à integração destas duas áreas ainda terão de ser realizadas até que um sistema com o objetivo descrito neste artigo seja aplicado em um contexto amplo.

Para podermos concentrar no desenvolvimento do SAARME e reduzir o leque de problemas a serem enfrentados, lidamos com um campo bem restrito: indução do parto. Mesmo assim, podemos ter uma visão das limitações do sistema. A aplicação do SAARME em outras áreas da medicina requer profundos aperfeiçoamentos e não apenas correções do sistema. Deste modo, os resultados apresentados na tabela 1 não podem ser encarados como valores esperados de desempenho do SAARME em outros contextos. Certamente, dadas as limitações do sistema, o desempenho do SAARME será menor.

O desenvolvimento do SAARME foi realizado com amostras de resumos compreendendo o período 1986-1990, e envolvendo uma área bem restrita. Ele deve ser complementado com outros estudos que lidem com amostras mais recentes e de outras áreas da medicina. Também é importante que a análise dos resumos seja feita por mais de um pesquisador, de maneira cega, de modo a testar a consistência dos resultados obtidos.

Conforme mostrado na seção anterior, quando testado no grupo B, o programa SAARME produziu resultados corretos em 32 dos 83 títulos. Os resultados incorretos são devidos ao vocabulário desconhecido, estruturas não gramaticais ou à falha do sistema em reconhecer o modo como as informações são transmitidas. O vocabulário desconhecido pode ser assim agrupado: 1) palavras que possuem a mesma raiz de outras palavras no dicionário; 2) nomes de medicamentos e hormônios; 3) palavras que expressam estado, doenças, sítios anatômicos; 4) nomes de animais ou termos relacionados a eles; 5) nomes de tecnologias; 6) palavras usadas para indicar posição; e 7) outras palavras. Com exceção do grupo 7, os outros grupos são compostos de palavras que deveriam ser inseridas em uma versão mais completa do sistema. Entretanto, como o número de palavras de uma linguagem está continuamente crescendo, é necessário que mecanismos sejam implementados para lidar com algumas palavras desconhecidas que eventualmente possam aparecer.

Aumentando o vocabulário do sistema com as palavras desconhecidas, duas outras classes de problemas ainda dificultam a correta análise da sentença. Estruturas sintáticas não previstas na gramática do sistema impediram a identificação correta das tecnologias em 15 títulos. Isto reflete o estágio presente de SAARME, que pode ser aperfeiçoado com novas estruturas gramaticais analisáveis, mas também revela as deficiências do sistema ao lidar com situações não previstas. Finalmente o sistema identifica somente alguns padrões como as informações são transmitidas. Embora seja possível adicionar mais padrões, a recuperação de informações deve ter uma base mais racional, uma vez que há virtualmente infinitas formas de transmitir a mesma informação. Por outro lado, a redação mais objetiva e concisa dos títulos e resumos de trabalhos científicos pode facilitar o desenvolvimento de um sistema automático de leitura.

O uso de classes semânticas para dirigir a análise do texto parece proporcionar bons resultados e os objetos propostos para mapear as informações sobre tecnologias são organizados em uma estrutura de grafos que permite a recuperação posterior, caso haja necessidade. Este esquema de representação pode ser bastante útil em desenvolvimentos futuros.

A estratégia de análise do SAARME segue a sequência: análise sintática seguida da análise semântica. Apesar de ser uma estratégia comum (Ledorary *et alii*, 1992; Sager *et alii*, 1987), outros enfoques tais como a realização da análise semântica diretamente (Rassinoux *et alii*, 1994) ou o uso de algoritmos paralelos podem ser mais eficientes. Por um lado, eles evitam que uma análise semântica não seja realizada devida a sentenças gramaticalmente mal-formadas; por outro lado, eles geram mecanismos mais robustos e gastam menos tempo para realizar a análise das sentenças.

O SAARME se limitou à leitura dos títulos dos resumos. A leitura de todo o corpo do resumo pode proporcionar uma informação mais precisa do trabalho: tecnologias avaliadas, delineamento, indicações e condições de uso das tecnologias, características dos grupos de pacientes analisados e resultados. A análise desta amostra de resumos mostrou que, embora seja possível extrair na maior parte dos casos informações importantes para o processo de avaliação tecnológica, elas nem sempre são apresentadas de modo objetivo. A elaboração dos resumos na forma estruturada facilita a leitura do mesmo como também a análise automática. Nesta amostra, um pequeno número de resumos estava na forma estruturada. Uma análise superficial de uma amostra mais recente verificou que o número de resumos estruturados têm crescido, em parte devido às exigências de revistas especializadas (Ad Hoc Working Group, 1987). Este esforço deve ser continuado. Também é desejável que maiores cuidados sejam prestados no processo de revisão dos trabalhos de modo a corrigir erros de ortografia e sintaxe, pelo menos os mais óbvios.

A análise linguística do corpo dos resumos para extrair informações sobre avaliação tecnológica para toda a área da medicina terá que resolver problemas consideráveis, tais como níveis de ambiguidade, extensão do vocabulário e estruturas linguísticas, representação do conhecimento, elipses, referências, estruturas mal formadas, coesão das sentenças e outros. Todo tipo de informação que pode ser útil para resolver estes problemas deve ser utilizada. Neste aspecto, além dos autores, títulos, fontes e resumos, as bases de dados eletrônicas apresentam, para cada estudo, outros campos contendo termos que transmitem alguma informação relativa ao estudo. O MEDLINE, por exemplo, possui o campo dos descritores (MeSH - *Medical Subject Headings*), "*checktags*", campo que informa se o estudo é realizado em animais ou em seres humanos, se os pacientes são do sexo masculino ou feminino, se o estudo é comparativo, etc. Estes campos podem ser úteis na análise linguística e podem proporcionar um meio mais eficiente e confiável de recuperar as informações (Cimino e Barnett, 1993). Por exemplo, a análise dos descritores permite ao sistema restringir a busca para as tecnologias avaliadas somente para aquelas listadas nos descritores e também restringir o domínio da medicina que ele deve considerar na análise. Além disso, se um campo informa que o estudo foi realizado em animais, o sistema não deve considerar este estudo como uma avaliação de tecnologia. Assim, uma integração da análise linguística com a análise destes campos pode tornar o sistema mais eficiente.

Também é importante para o processo de construir uma base de dados orientada para tecnologia que as informações sejam padronizadas, de tal forma que elas sejam compatíveis com nomenclaturas tradicionais e esquemas de classificação. Seguindo esta linha, uma análise da terminologia médica e a busca de linguagem para representação de conceitos médicos devem ser enfatizadas (Evans et alli, 1994).

CONCLUSÃO

Este artigo apresenta um sistema que analisa títulos de resumos e informa que tecnologias são avaliadas no estudo. Apesar de suas limitações, o SAARME produziu resultados corretos em 71,1% dos títulos no domínio restrito em que foi aplicado. Futuras implementações deste sistema devem considerar duas direções: análise do corpo dos resumos para extrair outros parâmetros de avaliação e integração de técnicas de interpretação da linguagem natural com técnicas de busca bibliográfica baseada em descritores. Dois fatores que podem contribuir para a implementação de um sistema automático são: a difusão da prática de redação dos resumos na forma estruturada e a busca de uma terminologia para representar os conceitos médicos. Apesar dos inúmeros problemas que precisam ser solucionados, uma base de dados orientada para tecnologias, montada de modo automático, se constituirá num excelente instrumento para a difusão e organização do conhecimento sobre avaliação tecnológica em saúde.

AGRADECIMENTOS

Os autores agradecem o financiamento do CNPq e a COPPE/UFRJ onde a maior parte deste trabalho foi desenvolvido.

AD HOC WORKING GROUP (1987). "A Proposal for More Informative Abstracts of Clinical Articles". Ad Hoc Working Group for Critical Appraisal of the Medical Literature. *Annals of Internal Medicine*. v. 106, p. 598-604.

CHALMERS, I. (ed)(1988). *Oxford Database of Perinatal Trials*. Oxford: Oxford University Press.

CHALMERS, I. and HAYNES, B. (1994). "Reporting, Updating, and Correcting Systematic Reviews of the Effects of Health Care". *British Medical Journal*. v. 309, p. 862-865.

CHASSIN, M. R., PARK, R. E., FINK, A., RAUCHMAN, S., KEESEY, J. and BROOK, R.H. (1986). *Indication for Selected Medical and Surgical Procedures. A Literature Review and Ratings of Appropriateness. Coronary Arterial Bypass Graft Surgery.* Rand Corporation, R-3204/2-CWF/HF/HCFA/PMT/RNJ.

CIMINO, J. J. and BARNETT, G. O. (1993). "Automatic Knowledge Acquisition from MEDLINE", *Methods of Information in Medicine*. v. 32, n. 2, p. 120-130.

EVANS, D. A., HERSH, W. R., MONARCH, I. A., LEFFERTS, R. G. and HANDERSON, S. K. (1991). "Automatic Indexing of Abstracts Via Natural Language Processing Using a Simple Thesaurus". *Medical Decision Making*. v. 11, suplemento, p. S108-S115.

EVANS, D. A., CIMINO, J. J., HERSH, W. R., HUFF, S. M. and BELL, D. S. (1994). "Toward a Medical-concept Representation Language". *Journal of the American Medical Informatics Association*. v. 1, n. 3, p. 207-217.

FREIRE, S. M. (1994). *Extração Automática de Informações Relativas a Tecnologias em Saúde a Partir dos Resumos Publicados na Literatura Médica*. Tese de Doutorado, Programa de Engenharia Biomédica, Rio de Janeiro: COPPE/UFRJ, 238 p., set.

LEDORAY, V., GIUSIANO, B. and ROUX, M. (1992). "A system for understanding Medical Reports: Architecture and Knowledge Required". MEDINFO92, p. 1389-1394.

MARCUS, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. Cambridge: MIT Press.

RASSINOUX, A.M., MICHEL, P.A., JUGE, C., BAUD, R. and SCHERRER, J.R. (1994). "Natural Language Processing of Medical Texts within the HELIOS Environment". *Computer Methods and Programs in Biomedicine*. v. 45, suplement, p. S79-S96.

SAGER, N., FRIEDMAN, C. and LYMAN, M. S. (1987). *Medical Language Processing: Computer Management of Narrative Data*. Reading: Addison-Wesley.

NATURAL LANGUAGE PROCESSING APPLIED TO THE AUTOMATIC EXTRACTION OF HEALTH TECHNOLOGY ASSESSMENT INFORMATION

S. M. Freire[1] and R. B. Panerai[2]

ABSTRACT -- The aim of this work is to evaluate the possibility of developing a computational system to retrieve information on health technologies from abstracts of papers published in the medical literature. A sample of abstratcts of papers dealing with cervical ripening and/or labor induction was collected from MEDLINE and divided into two groups: one group was used for analysis and implementation of the system. The other group was used to test the capability of the system to retrieve the evaluated technologies from the title of each study. The program produced correct information in 71.1% of the abstracts. This study shows the potential of using natural language processing techniques to retrieve health technology information from medical abstracts and build a technology-oriented database system.

Keywords: Natural Language Processing, Information Retrieval, Medical Literature

[1] Programa de Engenharia Biomédica - COPPE/UFRJ and FUNREI - Fundação de Ensino Superior de São João del Rei, MG, Brazil
Current Position: Associate Professor - Medical Informatics/LAMPADA/FCM/UERJ, Rua Prof. Manuel de Abreu 48/2^0 andar, CEP 20560-001, Vila Isabel, Rio de Janeiro, RJ, Brazil
E-mail: sfreire@saude.lampada.uerj.br

[2] Programa de Engenharia Biomédica - COPPE/UFRJ
Current Position: Senior Lecturer, Department of Medical Physics, Leicester University, LE1 5WW, Leicester, England, E-mail: rp9@leicester.ac.uk