

Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde

DOI: 10.3395/reciis.v3i1.244pt



Lucelene Lopes

Universidade Católica do
Rio Grande do Sul, Porto
Alegre, Brasil
lucelene.lopes@pucrs.br



Renata Vieira

Universidade Católica do
Rio Grande do Sul, Porto
Alegre, Brasil
renata.vieira@gmail.com

Maria José Finatto

Universidade Federal do Rio Grande do Sul, Porto
Alegre, Brasil
mfinatto@pq.cnpq.br

Daniel Martins

Universidade Católica do Rio Grande do Sul, Porto
Alegre, Brasil

Adriano Zanette

Universidade Federal do Rio Grande do Sul, Porto
Alegre, Brasil

Luiz Carlos Ribeiro Jr

Universidade Católica do Rio Grande do Sul, Porto
Alegre, Brasil

Resumo

Neste artigo mostramos o uso da ferramenta OntoLP no processo de construção de ontologias em um experimento na área da Saúde. Especificamente, faz-se a extração de termos com base em um corpus da área de Pediatria. Comparamos o resultado obtido pela ferramenta com os resultados de referência de uma lista de termos obtida manualmente. Nessa comparação, são analisados bi-gramas e tri-gramas obtidos através de diferentes métodos. Concluímos o trabalho observando as vantagens do processamento com inclusão de informação lingüística complexa, como análise sintática e semântica.

Palavras-chave

processamento de linguagem natural; ontologias; construção de ontologias para área da saúde; extração semi-automática de termos

Introdução

O desenvolvimento da informatização e também a evolução constante dos meios para armazenamento de grandes massas de dados, nos mais diversos setores da sociedade, deram origem a uma significativa quantidade de bases digitais e fontes de dados. A maior parte dessas disponíveis na internet, e das mais variadas espécies, tais como textos, imagens, vídeos, serviços, hipertextos etc. Nesse sentido, faz-se necessário aprimorar modelos descritivos do conhecimento disponibilizado por esses recursos para que sejam recuperados quando necessários. A falta de padronização na representação de conhecimento pode dificultar a compreensão do conteúdo das diversas bases, inviabilizando, conseqüentemente o seu uso. Como uma alternativa para resolver esse problema, a área de sistemas de informação e ciência da computação tem adotado o uso de representação de conhecimento por ontologias (Gomez-Perez et al. 2004).

Ontologias têm sido empregadas como forma de conceituar, estruturar e representar, em um documento, o conhecimento de um domínio de forma que possa ser compartilhado. Esta prática tem sido adotada em vários domínios, e em especial na Biologia, Bioinformática, Biomedicina e na Medicina, que é o domínio de conhecimento explorado nesse artigo. Entretanto, é sabido que ontologias têm um processo de construção trabalhoso, que exige muito tempo e esforço, principalmente na sua utilização em larga escala (Brewster et al. 2003). Uma solução para isso é investir em pesquisas para que se consiga automatizar a tarefa de construção de ontologia de domínios específicos (Buitelaar et al. 2003). Essas pesquisas consideram, muitas vezes, as bases de texto como fontes de conhecimento. Essas fontes, por sua vez, estão expressas em diferentes idiomas, fazendo com que métodos baseados no uso de informações lingüísticas sejam desenvolvidos para as diferentes línguas.

Nesse contexto, apresentamos um estudo voltado especificamente para a língua portuguesa considerando o domínio da Medicina, em particular a área de Pediatria. Neste estudo destacamos a comparação de alternativas de abordagem para identificação de termos compostos (conceitos expressos em mais de uma palavra), considerando que esta é apenas a etapa inicial no processo de construção de ontologias.

Na área de Medicina, mais especificamente no tema da prevenção e promoção de Saúde, por exemplo, o tratamento automatizado da informação textual tende a ajudar pesquisadores e gestores de políticas de informação a reconhecer os melhores modos de apresentar os dados mais relevantes em função dos objetivos e da situação comunicativa que se tenha. Desenha-se aqui a área de pesquisa já reconhecida, fora do Brasil, como *e-health*. A idéia de compactação e da representação de extratos da informação em ciências da saúde (e seus desafios) tem tido espaço para discussão não só no âmbito do reconhecimento de linguagens e de terminologias científicas que precisam ser “facilitadas” para o leigo, mas também nas próprias publicações médicas. Para tanto, é preciso que sejam desenvolvidos sistemas de processamento da in-

formação e esse desenvolvimento precisa estar a cargo de equipes multidisciplinares compostas por profissionais da saúde, estudiosos da língua e da comunicação em ciências e engenheiros que constroem sistemas informatizados com interfaces práticas e de uso simples.

O Journal of Medical Internet Research – JMIR <<http://www.jmir.org/>>, por exemplo, contempla temas de *e-health*. Os resumos, extratos de texto ou representações de conteúdo são gerados com apoio de softwares ou de programas e têm saída em forma de:

- 1) texto sintético recortado a partir do todo de um texto fonte ou de grupos de textos;
- 2) esquemas do tipo mapa conceitual a partir de um texto ou grupos de textos fonte;
- 3) esquemas de relações hierárquicas de nódulos conceituais em ontologias a partir de um ou de mais textos.

Em meio ao cenário brasileiro, podemos pinçar um exemplo ilustrativo da utilidade de sistemas de *e-health*. No caso do Ministério da Saúde, dada a propaganda institucional veiculada a respeito no final de 2008 e início de 2009, sabemos que, por exemplo, a hanseníase ainda é uma doença de considerável impacto no Brasil. Entretanto, apesar da mobilização do Ministério e da sociedade civil organizada, a população ainda parece refratária às campanhas informativas que divulgam medidas preventivas. Entre vários recursos, as matérias sobre prevenção da hanseníase veiculadas na televisão e mídias impressas parecem não surtir um efeito desejado.

Nesse caso particular, um mapeamento da informação disponível, sobretudo em veículos de popularização de temas de saúde, feito a partir de em grandes massas de dados textuais *on-line*, de textos científicos a textos científicos dirigidos a leigos, pode vir a mostrar, por exemplo, que a palavra “lepra”, por mais carregada negativamente ou estigmatizante que seja, raramente é mencionada nos textos de divulgação que tratam sobre hanseníase. Talvez essa lacuna de inter-relação possa justificar o não entendimento mais imediato das mensagens dirigidas pelo público leigo. Um mapeamento semelhante poderia também mostrar situações de usos de termos conexos e de seus equivalentes mais ou menos populares em textos dirigidos a especialistas ou a semi-especialistas e se haveria, inclusive, alguma ponderação mais ou menos presente a respeito do emprego dessas terminologias, em seus diferentes matizes, por parte da própria comunidade científica.

Assim, seria possível que o gestor percebesse, a partir de um levantamento estatístico da configuração da linguagem em textos que visam informar os leigos, que a linguagem empregada precisaria favorecer *links* nocionais ou lógicos entre um nódulo conceitual X e um nódulo conceitual Y (com a devida informação sobre o caráter das diferentes denominações). Diagnosticada a falta e suprida a informação, um cidadão leigo, como no caso acima, poderia relacionar noções e, acionando sua memória cultural sobre uma dada doença ou risco, poderia colocar-se de modo pró-ativo perante a informação que recebe.

Um outro exemplo de utilidade desses sistemas, mas em outra dimensão, pode revelar como está sendo empregado um termo como “prevalência” pela comunidade de

médicos-pesquisadores em publicações científicas em um intervalo, por exemplo, de 2 anos de publicações ao longo de um corpus de mais de um milhão de palavras. Em um acervo de textos de Pediatria, publicados pelo Jornal de Pediatria <<http://www.jped.com.br>>, vê-se como o emprego dessa expressão pode gerar mal-entendidos quando tende a extrapolar a seara do termo técnico e funde-se à palavra da língua comum, cujo sentido é “o que predomina”, enquanto que, na terminologia médica, trata-se de um termo que corresponde a uma medida estatística em Epidemiologia. Nesses textos, a presença de uma construção como “prevalência de uso de chupeta” pode sinalizar uma informação importante para o editores da revista. Além de casos específicos sobre o emprego de termos técnicos, há outros casos indicados de modo esparso pela própria comunidade médica. São situações que extrapolam as terminologias e alcançam o estatuto de construções recorrentes nos textos, tal como vemos no trabalho *Expressões médicas: falhas e acertos* (Bacelar et al. 2003).

Para além da observação de usos de linguagem, a prospecção da informação científica disponível em grandes acervos de periódicos científicos é importante também porque “o aumento da produção de conhecimento - e, portanto, do número de periódicos - na segunda metade do século passado levou a comunidade de profissionais e pesquisadores a encarar o desafio de desenvolver critérios de qualidade que pudessem orientar os leitores na seleção da melhor evidência científica” (Blank et al. 2006, p.97). Esses tipos de dados, ao serem obtidos e estudados extensivamente em grandes bases de textos disponíveis na internet, oferecem informações valiosas para os gestores e editores de periódicos especializados à medida que possibilitam um retrato da recorrência de temas e de noções de uma dada comunidade científica. Essa comunidade teria, então, a partir das técnicas de exploração e de representação informatizada de bases de conhecimento, acesso a um amplo quadro de informações sobre a sua própria prática de comunicação, o que pode ser útil para ponderar sobre a feição e a circulação do conhecimento produzido.

Os exemplos citados, relativos a simples incidências de palavras em textos, visam ilustrar as possibilidades de uso de sistemas avançados de tratamento de informação textual. Esses sistemas, além da localização da informação propriamente dita, mostram como a informação está configurada lingüística e nocionalmente em diferentes situações. Revelam também que outras unidades de vocabulário acompanham-nas de modo mais freqüente ou raro. São sistemas inovadores que integram buscadores e ordenadores de dados em acervos textuais, que integram estatística lexical, sumarização (sintetização de conteúdos) e ontologias. São ferramentas de seleção de informação que precisam de metodologias qualificadas para o tratamento dos fenômenos da linguagem científica em uso. Nesses sistemas, a cooperação entre profissionais de saúde, de lingüistas e cientistas de computação é uma necessidade.

Neste artigo é apresentada a ferramenta OntoLP (Ribeiro 2008), que visa auxiliar de forma semi-automática os engenheiros de ontologias de língua portuguesa, sejam

eles especialistas do domínio em questão (profissionais da saúde) ou lingüistas. A ferramenta mostra sugestões de termos, conceitos e de organização de hierarquias da ontologia, com base no conhecimento registrado em uma base textual ou corpus de domínio.

Mais especificamente, o artigo apresenta um estudo sobre a análise e identificação de termos compostos, ou seja, termos que contêm duas (bi-gramas) ou mais palavras (n-gramas). No contexto deste trabalho, apenas bi-gramas e tri-gramas são extraídos e isso corresponde à etapa inicial no complexo processo de construção de ontologias. Métodos alternativos de processamento textual são comparados. O estudo é desenvolvido através de uma aplicação na área da saúde, considerando um corpus de Pediatria e uma lista de termos de referência para avaliação dos métodos.

Ferramenta OntoLP

O processamento de linguagem natural (PLN) se faz presente na construção de ontologias através de textos. O PLN utiliza técnicas lingüísticas baseadas em análises sintáticas e morfológicas dos textos representando as informações em vários níveis.

Como as ontologias criadas ficam restritas a um idioma específico, a dificuldade aumenta no domínio do Português, língua sobre a qual pouco foi feito se comparada com o Inglês. Em resposta a essa dificuldade, novas metodologias para construção semi-automática de ontologias vêm sendo criadas juntamente com as ferramentas de auxílio para essa construção.

OntoLP é uma ferramenta, na verdade um *plug-in* para o editor de ontologias Protégé (Gennari et al. 2002), um editor bastante utilizado na comunidade científica e que dá suporte à construção de ontologias, seguindo as tecnologias da Web Semântica, como por exemplo, a construção de ontologias OWL *Web Ontology Language*, conforme o padrão definido pelo *World Wide Web Consortium* (W3C) (McGuinness et al. 2004).

O processo de construção automática de ontologias é dividido em cinco etapas básicas (Buitelaar 2005): extração de termos candidatos a conceitos de um domínio; identificação da relação hierárquica entre os termos; identificação de relações não-hierárquicas; identificação de instâncias e extração de regras (axiomas). Esse processo pode ser representado em camadas de acordo com a representação feita na Figura 1.

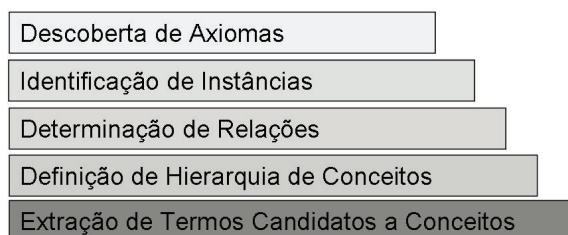


Figura 1 - Etapas básicas de construção de ontologias.

Do ponto de vista deste trabalho, que tem por objetivo em longo prazo a construção automática de ontologias, a extração de termos é a tarefa inicial e fundamental, pois os termos extraídos representam os conceitos de uma área específica e são a base para a execução das demais fases.

Para extração de termos, existem três abordagens principais:

- Estatística – os documentos contidos no corpus são vistos como um conjunto de termos e é medida sua frequência de ocorrência;
- Lingüística – os textos são anotados com informações lingüísticas (morfológicas, sintáticas e semânticas) e estas informações são levadas em consideração no processo de extração;
- Híbrida – considera a união das duas abordagens (estatística e lingüística).

A ferramenta OntoLP (Ribeiro 2008) é composta de uma série de métodos híbridos. Estes métodos de extração de termos estão agrupados em duas etapas:

- CorpusXCES: nesta etapa é realizada a anotação do corpus com informações lingüísticas pelo *parser* PALAVRAS (Bick 2000). O corpus anotado contém informações morfológicas, sintáticas e semânticas, representadas no formato XCES/PLN-BR (Ide et al. 2000). O processamento textual para extração de termos é baseado neste corpus anotado. Através da análise morfossintática, informações são adicionadas ao texto original que permitem explorar o emprego de métodos menos ou mais informados linguisticamente. Nesse trabalho, as informa-

ções lingüísticas empregadas pelos métodos de extração são as categorias gramaticais das palavras (por exemplo, *substantivo, verbos, adjetivos*), categorias semânticas prototípicas (por exemplo, *humanos, animais, doenças*) e a identificação de grupos gramaticais nominais (sintagmas nominais como *aleitamento materno exclusivo*);

- Extração de termos: para essa etapa, são aplicados diferentes métodos que combinam medidas estatísticas de frequência com as informações lingüísticas mencionadas acima, com a finalidade de extrair os termos simples (uni-gramas) e de termos complexos (n-gramas, onde $n > 1$).

No *plug-in* OntoLP, os métodos de extração possuem um conjunto de funcionalidades para auxiliar o engenheiro de ontologias nas etapas que podem ter interação humana. A interface de extração divide-se em três partes: (1) seleção de grupos semânticos; (2) extração de termos simples; e (3) extração de termos compostos.

A Figura 2 mostra de maneira geral como essas funcionalidades aparecem na ferramenta OntoLP, onde os números 1, 2 e 3 indicam as interfaces de extração citadas acima. Na figura, o grupo semântico selecionado é (H - Humano). A janela mais à direita mostra as palavras do grupo em forma de *tag clouds*, onde os termos mais frequentes são destacados, no caso <paciente> e <criança> aparecem como as mais frequentes. O engenheiro de ontologias pode optar por excluir ou não um grupo semântico. Nesse exemplo, o grupo H (humano) não seria excluído devido à sua relevância no corpus analisado.

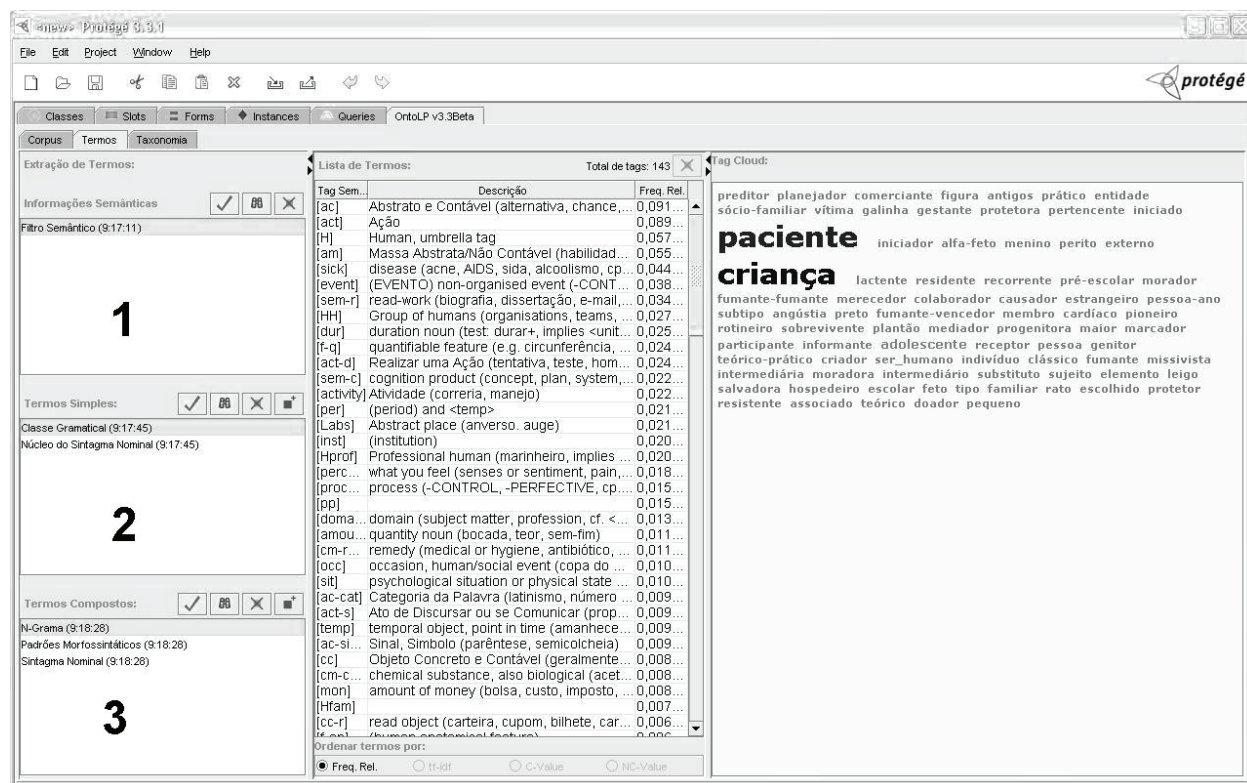


Figura 2 - Interface de extração de termos e suas etapas propostas para tarefa.

A etapa de Seleção de Grupos Semânticos é opcional. A ferramenta mostra ao usuário as informações semânticas que o *parser* (etiquetador) PALAVRAS associa às palavras do corpus. São informações prototípicas que classificam nomes comuns em classes gerais, por exemplo,

a tag “<an>” atribuída ao substantivo “músculo”, indica que a palavra pertence à classe “Anatomia”. A Figura 3 mostra alguns exemplos destes grupos e subgrupos semânticos existentes no corpus de Pediatria analisado neste artigo.

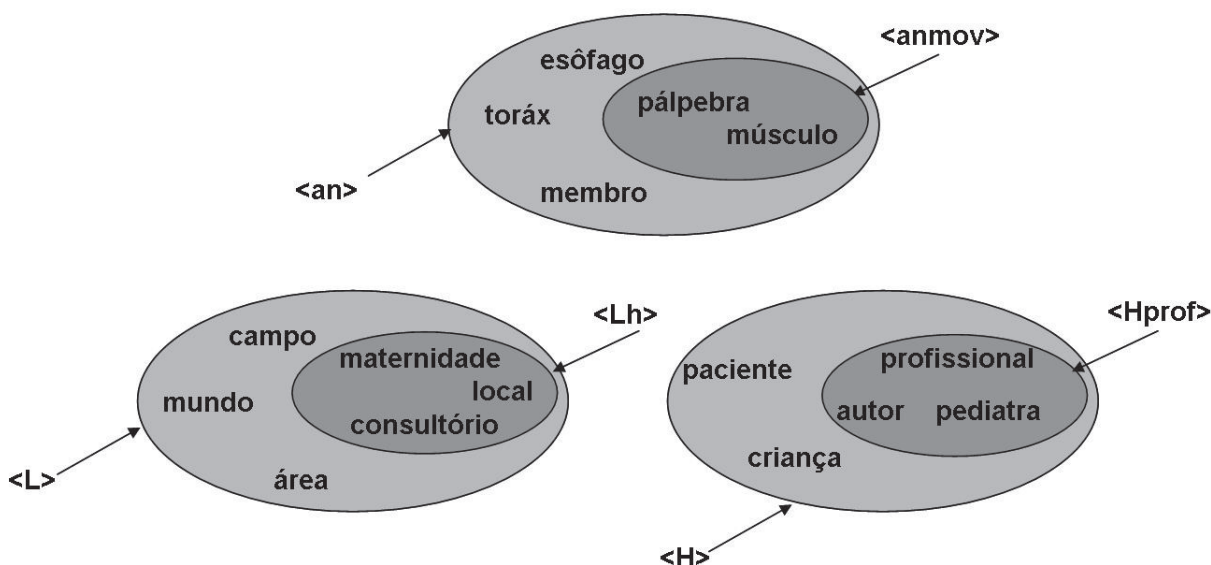


Figura 3 - Exemplos de grupos e subgrupos semânticos.

Dessa forma, os substantivos etiquetados com uma mesma *tag* (etiqueta) são agrupados em conjuntos semânticos, por exemplo:

- Grupo <an> (Anatomia): {esôfago, tórax, membro, pálpebra, músculo}
- Grupo <L> (Lugar): {campo, mundo, área, maternidade, local, consultório}
- Grupo <H> (Humano): {paciente, criança, profissional, autor, pediatra}

Os grupos semânticos podem ainda apresentar subdivisões, como pode ser observado na Figura 3: a) anatomia (<an>) e anatomia de movimento (<anmov>); b) lugar (<L>) e lugares funcionais (<Lh>); c) humano (<H>) e humano profissional (<Hprof>).

A ferramenta OntoLP disponibiliza ao usuário o método Filtro por Grupo Semântico, que emprega os seguintes passos:

- As *tags* semânticas presentes no corpus de entrada são extraídas;
- O cálculo de frequência relativa (FR) é aplicado a listas de *tags* semânticas que são apresentadas ao engenheiro ordenadas conforme essa medida;
- O engenheiro exclui os grupos semânticos que considera não ter relação com o domínio representado pelo corpus de entrada.

Esse método pode ser considerado como a construção de uma lista de *stopwords* específicas para um domínio. Essas *stopwords* são itens a não considerar. A

seleção correta dos grupos depende do conhecimento do engenheiro de ontologia sobre a área de conhecimento implicada. A ferramenta auxilia o engenheiro ao mostrar as ocorrências dos termos de cada grupo e sua relevância pelo método de *tags clouds*, ou seja, um método que atribui fontes maiores e destaque textual para termos mais frequentes no corpus.

Após a seleção de grupos semânticos, são executadas as segunda e terceira etapas da extração (extração de termos simples e extração de termos compostos), implementadas por métodos híbridos (estatístico e lingüístico). Na segunda etapa, é realizada a extração de termos simples. O método utilizado é o método de classes gramaticais, detalhado em (Ribeiro 2008).

A terceira etapa, extração de termos compostos, foco deste trabalho, consiste em identificar bi-gramas e tri-gramas. Nessa fase, utilizamos três diferentes métodos, com diferentes complexidades lingüísticas. Em primeiro lugar, é realizada a extração de n-gramas por frequência de ocorrência, simplesmente, com aplicação de filtros simples como eliminação de termos com preposições iniciais ou finais. O segundo método leva em consideração a classe gramatical dos termos e padrões de extração, tais como:

- substantivo adjetivo - aleitamento materno
- substantivo preposição substantivo - saturação de oxigênio

O terceiro método realiza a extração dos sintagmas nominais, tais como reconhecidos pelo analisador sintático. Este é um nível de informação lingüístico estrutural

mais complexo e a sua produção requer ferramentas especializadas.

No decorrer das etapas de extração de termos simples e compostos, os métodos recebem a lista de grupos semânticos gerada na primeira etapa, e percorrem o corpus selecionando os termos que fazem parte de pelo menos um grupo presente na lista de entrada. A ferramenta oferece como opção quatro medidas de relevância: FR, *tf-idf* (Manning et al. 1999), NC-Value e C-Value (Frantzi et al. 1998). Os termos extraídos são organizados de forma decrescente com base nos resultados da aplicação dessas medidas, podendo o engenheiro de ontologias analisar e editar a lista final de termos. Cabe ressaltar que as análises feitas neste artigo levaram em conta a Freqüência Relativa (FR), a qual considera o número de vezes que um termo aparece no documento dividido pelo total de palavras do documento.

Experimentos

O corpus utilizado nos experimentos com a ferramenta é composto por 283 textos em português extraídos

do Jornal de Pediatria, num total de 785.448 palavras. Sobre esse corpus foram realizados os experimentos que podem ser descritos conforme a Figura 4. Inicialmente o corpus foi anotado pelo *parser* PALAVRAS, que gerou uma representação XML, sendo convertida para arquivos tipo XCES. Esse corpus anotado em formato XCES foi lido pelo plug-in OntoLP. A partir disso, a extração de termos compostos seguiu as etapas:

- extração de grupos semânticos - através de uma análise manual, o especialista exclui os grupos semânticos considerados não relevantes, os grupos semânticos são criados através da etiquetagem semântica do PALAVRAS e são verificados pelo usuário por meio das *tag clouds*;
- extração dos termos simples; e
- extração dos termos compostos em que foram analisados os métodos: n-gramas, padrões morfossintáticos e o sintagma nominal.

Neste artigo, comparamos os resultados obtidos com e sem a extração de grupos semânticos. Na extração sem exclusão de grupos semânticos, todos os termos são considerados na computação das etapas posteriores.

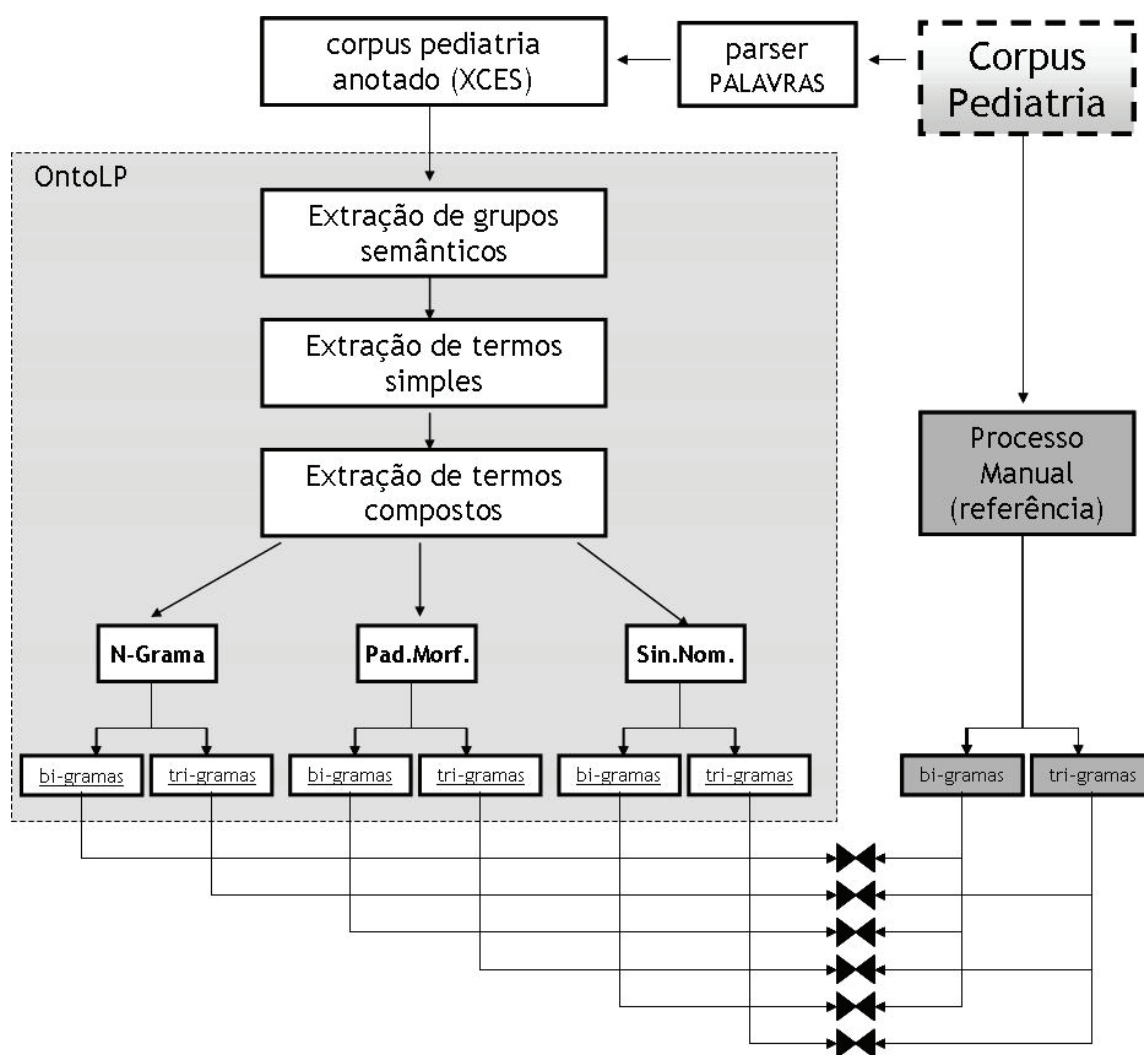


Figura 4 - Metodologia usada nos experimentos.

As etapas de extração de termos simples e compostos são reproduzidas para os dois grupos de termos identificados (com e sem exclusão de grupos semânticos) de maneira idêntica, tendo como saída do processo seis listas de bi-gramas e seis listas de tri-gramas. Cada uma delas foi comparada com as listas de referência de bi-gramas e tri-gramas. As listas de referências foram construídas por um processo fortemente apoiado em tarefas manuais, executado pelo Grupo TEXTQUIM/TERMISUL da Universidade Federal do Rio Grande do Sul (TEXTQUIM/UFRGS, <<http://www.ufrgs.br/textquim>>). O trabalho de extração de termos presentes no corpus de Pediatria visou à elaboração de um glossário para apoio aos estudantes de tradução. Esse material gerado também abastece os itens de um Catálogo de Expressões Recorrentes em Pediatria. O glossário e catálogo, desenhados como recursos *on-line* para educação à distância, visam auxiliar a formação de tradutores e de revisores de textos de Pediatria.

Na geração das listas de referência, foram inicialmente considerados n-gramas com mais de 5 ocorrências no corpus, extraídos automaticamente. A partir desta lista de 36.741 n-gramas, partiu-se para um processo de filtragem automático baseado em heurísticas. Por exemplo, termos que começavam ou terminavam por preposições foram transformados pela exclusão destas preposições; n-gramas contidos em n-gramas maiores foram excluídos. Dessa forma, um bi-grama que aparecia em um tri-grama foi descartado, pois, para fins de aprendizagem de tradução, termos com um maior número de palavras são preferíveis a termos menores. Por exemplo, o termo “aleitamento materno exclusivo” consta como um tri-grama, portanto “aleitamento materno” não consta na lista de bigramas. O processo resultou em uma lista com 3.645 n-gramas. Esta lista foi conferida manualmente por estudantes de tradução, resultando em uma lista final com 2.407 termos, sendo 1.293 bi-gramas, 775 tri-gramas e 339 termos de composição maior que 3 palavras.

A comparação das listas obtidas pela ferramenta OntoLP com as listas de referência foi feita através das seguintes métricas: precisão (P), abrangência (A) e *f-measure* (F). A precisão (P) indica a capacidade do método de identificar os termos corretos, considerando a lista de referência e é calculada pela fórmula (1).

$$P = (\text{Termos Referência} \cap \text{Termos Extraídos}) / \text{Termos Extraídos} \quad (1)$$

A abrangência (A) avalia a quantidade de termos corretos extraídos pelo método e é calculada através da fórmula (2).

$$A = (\text{Termos Referência} \cap \text{Termos Extraídos}) / \text{Termos Referência} \quad (2)$$

A *f-measure* (F) é a medida harmônica entre a precisão e abrangência, e é dada pela fórmula (3).

$$F = (2 * P * A) / (P + A) \quad (3)$$

A Tabela 1 apresenta o número total de termos encontrados nos experimentos para cada uma das análises feitas. Adicionalmente, são mostrados quantos desses termos estão presentes na lista de referência que possui um total de 1.293 bi-gramas e 775 tri-gramas. O número de termos recuperado é bem superior ao número de termos da referência, uma vez que todos os termos extraídos do corpus são considerados, sem a utilização de um ponto de corte por frequência. Obviamente, este número diminui com a exclusão de grupos semânticos. Nesse caso, como a proporção de termos não relevantes excluídos é maior do que a de termos relevantes, observa-se um aumento na precisão.

Tabela 1 - Termos extraídos e listas de referência

Com exclusão de grupos semânticos	n-gramas		Padrões morfosintáticos		Sintagma nominal	
	Bi-gramas <i>bi nG</i>	Tri-gramas <i>tri nG</i>	Bi-gramas <i>bi PM</i>	Tri-gramas <i>tri PM</i>	Bi-gramas <i>bi SN</i>	Tri-gramas <i>tri SN</i>
Total de Termos	13.115	18.554	27.763	27.322	8.926	5.959
Termos presentes na referência	610	407	636	406	588	283
Sem exclusão de grupos semânticos	n-gramas		Padrões morfosintáticos		Sintagma nominal	
	Bi-gramas <i>bi nG</i>	Tri-gramas <i>tri nG</i>	Bi-gramas <i>bi PM</i>	Tri-gramas <i>tri PM</i>	Bi-gramas <i>bi SN</i>	Tri-gramas <i>tri SN</i>
Total de Termos	18.325	23.588	33.276	30.497	12.691	7.509
Termos presentes na referência	769	451	780	441	740	311

As métricas finais da avaliação são apresentadas nas Figuras 5 e 6. O método que apresenta um melhor balanço entre precisão e abrangência (ambos para bi-gramas e tri-gramas) é a extração por sintagmas nominais, com exclusão de termos por grupos semânticos (f-measure de 11,51% e 8,41% para bi-gramas e tri-gramas, respectivamente). Estes são os métodos que empregam um maior nível de processamento lingüístico, tanto em relação ao

processamento sintático como processamento semântico. Esta observação indica que pré-processamento lingüístico do corpus tende a contribuir positivamente para a extração de termos compostos. Porém, ressaltamos que a informação semântica é empregada de forma semi-automática, os grupos são apresentados ao especialista do domínio que indica os grupos semânticos a serem desconsiderados no processo.

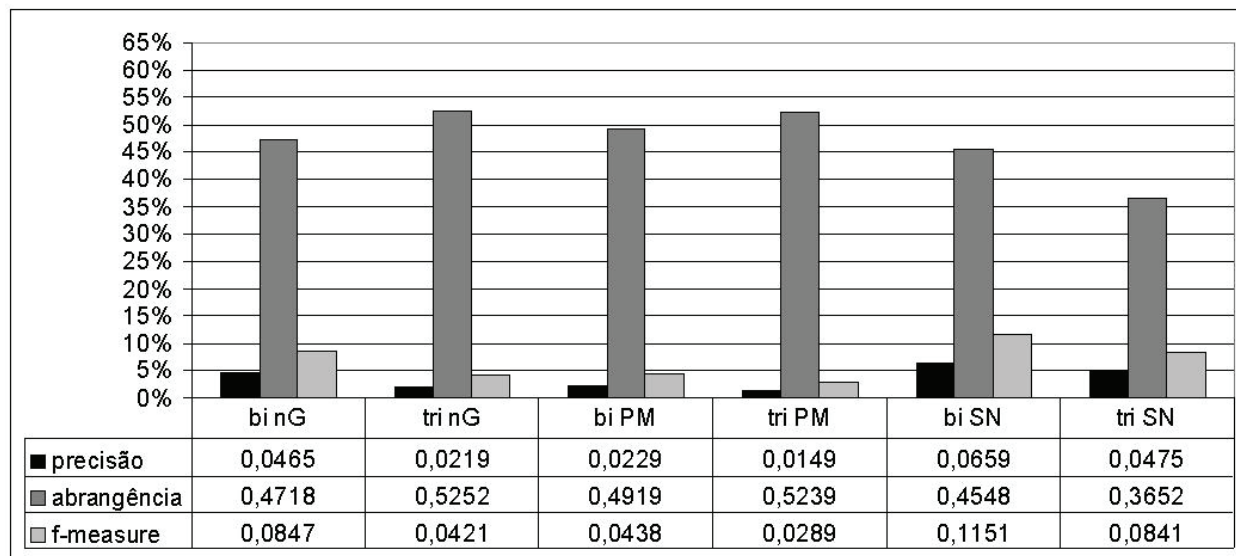


Figura 5 - Gráfico de métricas com exclusão de grupos semânticos.

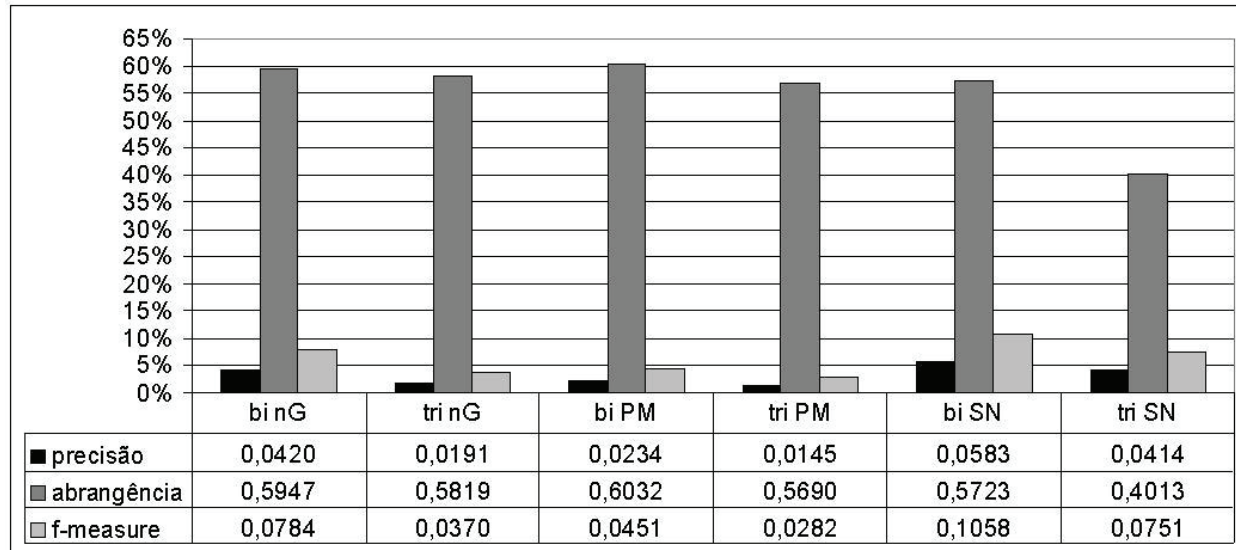


Figura 6 - Gráfico de métricas sem exclusão de grupos semânticos.

A abrangência atingiu apenas 61% dos termos da referência, ou seja, um número significativo dos termos da referência não foi encontrado por nenhum dos métodos empregados. Este aspecto ainda deve ser investigado.

Nos apêndices (1-4) estão listados os primeiros termos das listas extraídas, ordenados por frequência. Foram destacados em negrito os termos constantes da lista de referência. Pode-se observar que alguns termos

aparentemente relevantes, encontrados pelos métodos, não constam na referência, por exemplo, “idade gestacional”, “fator de risco” e “aleitamento materno”. Em alguns casos, bigramas relevantes não foram constatados na referência, pois esta não inclui sub-termos. O bi-grama “aleitamento materno”, por exemplo, está ausente da lista de referência, pois este é um sub-termo do o tri-grama “aleitamento materno exclusivo”. Fatos como

esse sugerem a necessidade de uma revisão do método de avaliação, por refinamento da referência, uma vez que os propósitos do presente trabalho diferem dos de auxílio à tradução. Por outro lado, destacamos a importância deste recurso, pois a área de aprendizagem de ontologias, por ser muito recente, ainda enfrenta sérios problemas relacionados à carência de recursos para avaliação. Avaliações, apesar de difíceis, são cruciais para o desenvolvimento das técnicas. Avaliações nesta área, além de lidarem com a escassez de recursos, enfrentam problemas relativos à subjetividade. É comum que diferentes especialistas tenham julgamentos diferenciados em relação à relevância dos termos.

Trabalhos relacionados

Em Cimiano (2006) aborda-se de forma bastante completa o problema de extração de ontologias de texto, um problema com muitas questões de pesquisa ainda em aberto.

Suchanek (2006) discute de forma geral o emprego de análise lingüística na extração de informação em bases textuais. Em particular, a extração de termos compostos é um tema bastante investigado (Ramisch et al. 2008) é um exemplo de trabalho recente nesta linha. O propósito de investigação de termos compostos, no entanto, é variável, nem sempre são considerados como parte do levantamento de conceitos em um domínio.

Trabalhos anteriores relativos ao português são poucos; (Baségio 2006), por exemplo, apresenta uma primeira abordagem para o problema de extração de ontologias de texto baseada em corpus de domínio. A abordagem de Baségio foi avaliada para o domínio do Turismo através do julgamento de especialistas, sem uma lista de referência. Nessa forma de avaliação, entretanto, não é possível calcular métricas como abrangência e *f-measure*.

Ribeiro e Vieira (2008) avaliam os métodos do Plug-in OntoLP num corpus de ecologia em relação aos primeiros 1000 termos extraídos por cada método. O impacto da extração de grupos semânticos é avaliado num conjunto de 150 termos. Neste trabalho apresentamos uma avaliação dos métodos a partir de uma lista de referência mais extensa, em outro domínio, a Pediatria, e as avaliações consideram o conjunto total de termos extraídos.

O fato de a área de aprendizagem de ontologias ser muito recente, torna-se difícil apresentar uma análise comparativa com outros trabalhos, uma vez que ainda não há testes padrões disponíveis.

Conclusão

Apresentamos aqui uma avaliação inicial do uso de técnicas de Processamento de Linguagem Natural aplicadas ao problema de construção de ontologias. Os experimentos realizados estão relacionados com a primeira etapa do complexo processo de construção de ontologias, qual seja, a etapa de identificação dos termos candidatos a conceitos. Os experimentos realizados consideram um corpus e uma lista de referência de construções relevantes para a área de ensino de tradução de textos de Medicina/Pediatria. Esta é

uma área, entre outras, que pode se beneficiar do desenvolvimento de técnicas de processamento textual e estruturação de informação, uma vez que há muito conhecimento específico adquirido, que está registrado em texto.

Apesar de preliminares, os resultados podem ser usados para observação do comportamento dos diferentes métodos de extração empregados. Os métodos informados linguisticamente mostraram vantagens em relação aos métodos menos informados.

Como trabalho futuro, pretendemos refinar tanto a lista de termos utilizada como referência como também o processo de extração de termos. Como o engenheiro de ontologias recebe uma lista de termos ordenada por frequência, uma avaliação importante a ser realizada é a análise da precisão dos termos mais frequentes; será importante avaliar também a evolução do balanço entre precisão e abrangência conforme o ponto de corte. Este trabalho já está em andamento. É importante identificar um balanço entre precisão/abrangência que seja útil, isto é, que possa contribuir positivamente ao engenheiro de ontologias.

Além disso, planejamos avançar nas outras etapas de construção de ontologias. Para isso, em um primeiro momento, trabalharemos com o agrupamento semântico das expressões, identificando hierarquias e similaridades entre os termos.

As técnicas avaliadas neste artigo são incorporadas ao editor de ontologias Protégé, por meio de um plug-in. O plug-in, bem como outros recursos para desenvolvimento de pesquisa em ontologias, está disponível em <<http://www.inf.pucrs.br/~ontolp>>.

Agradecimentos

Agradecemos aos financiamentos da Capes, CNPq e SEAD/UFRGS concedidos aos autores deste trabalho.

Referências bibliográficas

- Bacelar S, Galvão CC, Alves E, Tubino P. Expressões médicas: falhas e acertos. Rev Bras Cirurgia Cardiovascular, São Paulo. Jul./Set., 2003; 18(3).
- Baségio T. Uma Abordagem Semi-Automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil. 2006. Dissertação (Mestrado em Ciência da Computação), Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS, Porto Alegre.
- Bick E. The parsing System "Palavras": Automatic grammatical analysis of portuguese in a constraint grammar framework. PhD thesis, Arhus University. 2000.
- Blank D, Rosa LO, Gurgel RQ, Goldani MZ. Produção brasileira de conhecimento no campo da saúde da criança e do adolescente. J Pediatria, Rio de Janeiro. 2006; 82(2): 97-102.
- Brewster C, Ciravegna F, Wilks Y. Background and foreground knowledge in dynamic ontology construction. In: SIGI, Proceedings of the Semantic Web Workshop, 2003, Toronto. Proceedings. Toronto: August, 2003.

Buitelaar P, Cimiano P, Magnini B. Ontology learning from text: An overview. In: Buitelaar P, Cimiano P, Magnini B. (editors). *Ontology learning from text: methods, evaluation and applications*, v. 123 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2005.

Cimiano P. *Ontology learning and population from text: Algorithms, evaluation and applications*. Heidelberg: Springer-Verlag, 2006.

Frantizi KT, Ananiadou S, Ichi Tsujii J. The c-value/nc-value method of automatic recognition for multi-word terms. In: *ECDL'98: Proceedings of the second european conference on research and advanced technology for digital libraries*, 1998, London. *Proceedings*. Heidelberg: Springer-Verlag, 1998, p. 585-604.

Gennari J et al. The evolution of protégé: an environment for knowledge-based systems development. Technical Report SMI-2002-0943. 2002.

Gomez-Perez A, Corcho O, Fernandez-Lopez M. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Heidelberg: Springer-Verlag, 2004.

Ide N, Bonhomme P, Romary L. Xces: An xml-based encoding standart for linguistic corpora. In: *Proceedings of the second international language resources and evaluation conference*, 2000. *Proceedings*. Paris: European Language Resources Association, 2000.

Manning CD, Schutze H. *Foundations of statistical natural language processing*. Cambridge, Massachusetts: The MIT Press, 1999.


McGuinness DL, Van Harmelen F. OWL web ontology language overview. World Wide Web Consortium (W3C) recommendation. <<http://www.w3.org/TR/owl-features>>. Acesso em: 1 Feb. 2004.

PROTÉGÉ <<http://protege.stanford.edu>> Acesso em: 25 ago. 2008.

Ramisch C, Schreiner P, Idiart M, Villavicencio A. An evaluation of methods for the extraction of multiword expressions. In: *LREC 2008 MWE workshop: towards a shared task on multiword expressions*, Marrakesh, 2008. *Proceedings*. Paris: European Language Resources Association, 2008.

Suchanek FM, Ifrim G, Andweikum G. Leila: Learning to extract information by linguistic analysis. In: *Proceedings of the 2nd workshop on ontology learning and population: bridging the gap between text and knowledge*, Sydney, Australia, 2006. *Proceedings*. Association for Computational Linguistics, 2006.

Ribeiro LC. *OntoLP: Construção semi-automática de ontologias a partir de textos da língua portuguesa*. Dissertação (Mestrado em Computação Aplicada), Universidade do Vale do Rio dos Sinos - UNISINOS, São Leopoldo. 2008.

Ribeiro LC, Vieira R. *ontolp: engenharia de ontologias em língua portuguesa*. In: *Anais do xxviii congresso da SBC - SEMISH - Seminário integrado de software e hardware*, Belém do Pará, 2008. *Anais*. Porto Alegre: Sociedade Brasileira de Computação, 2008. 

Sobre os autores

Lucelene Lopes

É doutoranda do Curso de Ciências da Computação da PUCRS desde 2008. Possui Mestrado em Tecnologia em Saúde pela PUCPR (2007). Graduada em Ciências com Habilitação Plena em Matemática pela UNIVALE (2000). Atua principalmente em Aprendizagem de Máquina (Inteligência Artificial) desde o mestrado e mais recentemente, com o início do doutorado, tem atuado em extração de termos dentro da área de Processamento de Linguagem Natural.

Renata Vieira

Possui título de PhD em Informática pela University of Edinburgh (1998). É professora da PUCRS onde atua em pesquisa e ensino na graduação e pós-graduação na área de inteligência computacional, com ênfase em processamento de linguagem natural, representação do conhecimento, ontologias, agentes e web semântica. Possui experiência em coordenação de projetos, é membro de comitês científicos das principais conferências internacionais da área de lingüística computacional e agentes inteligentes. Participa ativamente no desenvolvimento da área de processamento de linguagem natural no país.

Apêndice 1

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (bi-gramas) **com** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

Bigramas NG

aleitamento materno, idade gestacional, ventilação mecânica, faixa etária, terapia intensiva, hipertensão arterial, baixo peso, grupo controle, **diferença estatística**, perímetro cefálico, grande número, período neonatal, alto risco, massa óssea, **exame físico**, vitamina d, **nível sérico**, grande risco, amamentação exclusiva, carga viral, **infecção urinária**, **diferença significativa**, grande frequência, significância estatística, **baixa estatura**, cicatriz renal, insuficiência adrenal, **otite média**, choque séptico, saúde pública, **diagnóstico diferencial**, insuficiência respiratória, regressão logística, nível plasmático, **prática clínica**, evolução neurológico, solução salina, **quadro clínico**, ventilação pulmonar, via oral, fibrose cística, **idade inferior**, relaxamento muscular, primeiro mês, grande incidência, grande prevalência, **tubo endotraqueal**, **frequência respiratória**, anemia falciforme, dieta isento, escolaridade materna, **avaliação clínica**, obesidade infantil, desconforto respiratório, dor abdominal, **escore z**, disfunção miccional, perda auditiva, hipertensão pulmonar, grau I, **escore clínico**, **evolução clínica**, **deposição pulmonar**, pressão intracraniano, alta hospitalar, perímetro braquial, fase aguda, tempo médio, sexto mês, longo prazo

Bigramas PM

aleitamento materno, faixa etária, idade gestacional, ventilação mecânica, criança pequena, terapia intensiva, hipertensão arterial, **diferença significativa**, perímetro cefálico, cicatriz renal, **otite média**, **efeito colateral**, período neonatal, criança grande, paciente pediátrico, **nível sérico**, **exame físico**, **infecção urinária**, **manifestação clínica**, massa óssea, amamentação exclusiva, **efeito adverso**, carga viral, significância estatística, **quadro clínico**, choque séptico, solução salina, ensaio clínico, saúde pública, insuficiência respiratória, atividade física, **doença crônica**, regressão logística, nível plasmático, **prática clínica**, evolução neurológica, evento adverso, **avaliação clínica**, via oral, **frequência respiratória**, **escore clínico**, ventilação pulmonar, **idade inferior**, grau I, dor abdominal, relaxamento muscular, fibrose cística, **tubo endotraqueal**, **critério diagnóstico**, **cardiopatia congênito**, **infecção respiratória**, exame complementar, anemia falciforme, pressão intracraniano, **evidência científica**, escolaridade materna, perda auditiva, desconforto respiratório, hipertensão pulmonar, exame laboratorial, obesidade infantil, **evolução clínica**, tempo médio, infecção congênito, disfunção miccional, **fator prognóstico**, **volume corrente**, paciente portador, **alergia alimentar**, doença pulmonar

Bigramas SN

aleitamento materno, idade gestacional, ventilação mecânica, período neonatal, hipertensão arterial, perímetro cefálico, cicatriz renal, terapia intensiva, alto risco, criança pequena, baixo peso, **exame físico**, massa óssea, **diferença significativa**, significância estatística, amamentação exclusiva, paciente pediátrico, saúde pública, **infecção urinária**, choque séptico, criança grande, **baixa estatura**, último ano, solução salina, via oral, **tubo endotraqueal**, fibrose cística, anemia falciforme, grande frequência, **cardiopatia congênito**, obesidade infantil, **prática clínica**, **efeito colateral**, **doença crônica**, pressão intracraniana, escolaridade materna, infecção congênito, alta hospitalar, **alergia alimentar**, **manifestação clínica**, exame complementar, **otite média**, **efeito adverso**, **frequência cardíaca**, evolução neurológica, doença pulmonar, disfunção miccional, **diagnóstico diferencial**, dois paciente, **frequência respiratória**, criança estudada, dor abdominal, **aleitamento exclusivo**, regressão logística, **quadro clínico**, hipertensão pulmonar, faixa etária, modo geral, bom resultado, carga viral, orelha média, idade escolar, longo prazo, alto frequência, evento adverso, ventilação pulmonar, curta duração, tamanho amostral, presente trabalho, quatro paciente

Apêndice 2

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (tri-gramas) **com** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

Trigrama NG

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, peso de nascimento, dia de vida, **aleitamento materno exclusivo**, unidade de terapia, **intervalo de confiança**, nível de significância, qualidade de vida, país em desenvolvimento, serviço de saúde, semana de vida, diferença estatística significativa, hora de vida, critério de inclusão, velocidade de crescimento, **curva de crescimento**, coleta de dado, diferença estatística significante, problema de saúde, **tipo de parto**, média de idade, **taxa de mortalidade**, termo de consentimento, tempo de internação, **grupo de risco**, consumo de medicamento, produção de leite, confiança de 95, **dieta de exclusão**, trabalho de parto, número de paciente, déficit de atenção, **saturação de oxigênio**, estilo de vida, prevalência de asma, necessidade de ventilação, **terapia intensivo neonatal**, uso de antibiótico, densidade mineral óssea, criança com idade, **risco de infecção**, paciente com doença, **faixa etária pediátrico**, livre de doença, volume de leite, plasmático de vitamina, **modelo de regressão**, **transtorno de ansiedade**, uso em criança, índice de massa, uso de droga, **ingestão de cálcio**, **radiografia de tórax**, período de tempo, ventilação não invasivo, **síndrome de Down**, equipe de saúde, **ponto de corte**, **solução salina hipertônica**, uso de

oxigênio, uso de medicamento, análise de variância, uso de medicação, **ventilação pulmonar mecânica**, nível de escolaridade, selo de água

Trigrama PM

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, peso de nascimento, profissional de saúde, dia de vida, grupo de paciente, unidade de terapia, **intervalo de confiança**, qualidade de vida, nível de significância, país em desenvolvimento, grupo de criança, serviço de saúde, semana de vida, hora de vida, critério de inclusão, maioria do paciente, velocidade de crescimento, problema de saúde, coleta de dado, **curva de crescimento**, tamanho da amostra, **tipo de parto**, número de paciente, média de idade, **taxa de mortalidade**, termo de consentimento, maioria do caso, tempo de internação, **grupo de risco**, **ponto de corte**, paciente do grupo, consumo de medicamento, produção de leite, início do sintoma, **dieta de exclusão**, trabalho de parto, paciente com doença, vida da criança, déficit de atenção, uso de antibiótico, **saturação de oxigênio**, estilo de vida, criança com idade, aumento da pressão, necessidade de ventilação, duração do aleitamento, aplicação da vacina, volume de leite, prevalência de asma, saúde da criança, **risco de infecção**, maioria da criança, uso em criança, **modelo de regressão**, **transtorno de ansiedade**, idade da criança, criança do sexo, **ingestão de cálcio**, índice de massa, **radiografia de tórax**, período de tempo, uso de droga, **tempo de queixa**, uso de oxigênio, uso de medicamento, início na infância

Trigrama SN

fator de risco, **aleitamento materno exclusivo**, profissional de saúde, peso de nascimento, critério de inclusão, coleta de dado, serviço de saúde, **tipo de parto**, tempo de internação, país em desenvolvimento, **faixa etária pediátrico**, trabalho de parto, **terapia intensivo neonatal**, **ventilação pulmonar mecânica**, uso de antibiótico, produção de leite, **intervalo de confiança**, **saturação de oxigênio**, velocidade de crescimento, **otite média aguda**, terapia intensivo pediátrico, qualidade de vida, uso de oxigênio, sala de parto, **dieta de exclusão**, nível de significância, ventilação não invasiva, estilo de vida, equipe de saúde, consumo de medicamento, análise de variância, **radiografia de tórax**, acidente de transporte, **grupo de risco**, **farelo de trigo**, selo de água, termo de consentimento, **relaxamento muscular inadequado**, uso de medicamento, uso de medicação, **doença de base**, densidade mineral óssea, **suplemento de cálcio**, diferença estatística significativa, **taxa de mortalidade**, critério de exclusão, **vacina contra influenza**, **centro de referência**, **ansiedade de separação**, nível de linfócito, deficiência de vitamina, criança mais velha, **tempo de queixa**, **curva de crescimento**, esquizofrenia com início, escape de ar, situação de estresse, **síndrome de abstinência**, infecção respiratória aguda, **amostra de sangue**, **tubo de ventilação**, **hemorragia digestiva alta**, hipótese de nulidade, centro de saúde, **controle sem hepatopatia**, **risco de infecção**, **uso de chupeta**,

vacinação contra influenza, **aspiração de mecônio**, local de trabalho

Apêndice 3

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (bi-gramas) **sem** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

Bigramas NG

aleitamento materno, idade gestacional, leite materno, faixa etária, ventilação mecânica, presente estudo, terapia intensiva, hipertensão arterial, leite humano, baixo peso, primeiro ano, estado nutricional, grupo controle, **diferença estatística**, perímetro cefálico, um ano, massa óssea, período neonatal, grande número, alto risco, 1 ano, **escore z**, **exame físico**, vitamina d, **nível sérico**, grande risco, análise estatística, carga viral, amamentação exclusiva, **infecção urinária**, **diferença significativa**, grande parte, dieta isenta, grande frequência, significância estatística, cicatriz renal, insuficiência adrenal, **baixa estatura**, **otite média**, história familiar, vez grande, choque séptico, **quadro clínico**, lesão pulmonar, saúde pública, regressão logística, insuficiência respiratória, **diagnóstico diferencial**, nível plasmático, pressão intracraniano, **prática clínica**, evolução neurológica, teste t, população estudado, solução salina, **diagnóstico precoce**, ventilação pulmonar, via oral, medicamento não, primeiro mês, **idade inferior**, fibrose cística, corticosteroíde antenatal, **avaliação clínica**, relaxamento muscular, grande incidência, grande prevalência, anemia falciforme, resposta inflamatória, **tubo endotraqueal**

Bigramas PM

aleitamento materno, faixa etária, idade gestacional, leite materno, ventilação mecânica, criança pequena, terapia intensiva, hipertensão arterial, leite humano, estado nutricional, **diferença significativa**, perímetro cefálico, cicatriz renal, massa óssea, **efeito colateral**, alimento complementar, **otite média**, período neonatal, análise estatística, **nível sérico**, paciente pediátrico, criança grande, **exame físico**, **manifestação clínica**, **infecção urinária**, **efeito adverso**, amamentação exclusiva, carga viral, dieta isento, **quadro clínico**, significância estatístico, vez grande, ensaio clínico, história familiar, solução salina, choque séptico, lesão pulmonar, população estudado, pressão intracraniano, atividade física, saúde pública, insuficiência respiratória, **prática clínica**, **doença crônica**, regressão logística, nível plasmático, maioria do paciente, lesão cerebral, evolução neurológico, evento adverso, **avaliação clínica**, **frequência respiratória**, ventilação pulmonar, **diagnóstico precoce**, **escore clínico**, via oral, **idade inferior**, relaxamento muscular, fibrose cística, dor abdominal, grau I, **tubo endotraqueal**, **critério diagnóstico**, **cardiopatía congênito**, resposta inflamatório, anemia falciforme, **infecção respiratório**, exame complementar

Bigramas SN

aleitamento materno, presente estudo, idade gestacional, leite materno, ventilação mecânica, dois grupo, 6 mês, leite humano, 2 ano, período neonatal, hipertensão arterial, cicatriz renal, perímetro cefálico, terapia intensivo, alimento complementar, alto risco, tabela 2, massa óssea, 5 ano, estado nutricional, um ano, criança pequena, baixo peso, **exame físico**, 1 ano, população estudado, **diferença significativa**, dois ano, cinco ano, análise estatística, significância estatística, tabela 3, amamentação exclusiva, paciente pediátrico, último década, tabela 1, 24 hora, saúde pública, criança grande, pressão intracraniano, choque séptico, **infecção urinária**, **baixo estatura**, último ano, seis mês, **tubo endotraqueal**, solução salina, via oral, fibrose cística, anemia falciforme, grande frequência, corticosteroide antenatal, 12 mês, amostra estudada, faixa etário, **cardiopatía congênito**, 10 ano, **efeito colateral**, obesidade infantil, quatro mês, país desenvolvido, mãe adolescente, 6 ano, **prática clínica**, **doença crônica**, 30 minuto, escolaridade materno, infecção congênito, **efeito adverso**, alta hospitalar

Apêndice 4

Este Apêndice apresenta os 70 primeiros termos identificados em cada método (tri-gramas) **sem** exclusão de termos. Os termos extraídos que constam na referência estão destacados.

Trigrama NG

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, **leite de vaca**, peso de nascimento, profissional de saúde, dia de vida, **aleitamento materno exclusivo**, unidade de terapia, grupo de paciente, **intervalo de confiança**, nível de significância, qualidade de vida, país em desenvolvimento, semana de vida, serviço de saúde, diferença estatística significativa, hora de vida, **ponto de corte**, critério de inclusão, velocidade de crescimento, polissacarídeo de soja, coleta de dado, diferença estatística significante, **curva de crescimento**, problema de saúde, isento de leite, t de Student, **tipo de parto**, grupo de criança, **taxa de mortalidade**, média de idade, termo de consentimento, tempo de internação, **grupo de risco**, confiança de 95, consumo de medicamento, produção de leite, **dieta de exclusão**, trabalho de parto, número de paciente, casca de banana, déficit de atenção, **saturação de oxigênio**, estilo de vida, necessidade de ventilação, densidade mineral óssea, criança com idade, uso de antibiótico, volume de leite, **terapia intensivo neonatal**, **risco de infecção**, prevalência de asma, livre de doença, paciente com doença, **faixa etária pediátrico**, uso em criança, plasmático de vitamina, **transtorno de ansiedade**, **modelo de regressão**, fórmula de soja, índice de massa, **radiografia de tórax**, uso de droga, **ingestão de cálcio**, período de tempo, **síndrome de Down**, tempo de queixa

Trigrama PM

ano de idade, fator de risco, mês de vida, ano de vida, mês de idade, **leite de vaca**, peso de nascimento, profissional de saúde, dia de vida, grupo de paciente, unidade de terapia, **intervalo de confiança**, nível de significância, qualidade de vida, grupo de criança, país em desenvolvimento, serviço de saúde, semana de vida, **ponto de corte**, hora de vida, critério de inclusão, maioria do paciente, velocidade de crescimento, problema de saúde, **curva de crescimento**, coleta de dado, **tipo de parto**, tamanho da amostra, número de paciente, maioria da vez, **taxa de mortalidade**, média de idade, maioria do caso, termo de consentimento, **grupo de risco**, tempo de internação, filho de mãe, consumo de medicamento, paciente do grupo, produção de leite, paciente com doença, trabalho de parto, início do sintoma, **dieta de exclusão**, vida da criança, casca de banana, déficit de atenção, **saturação de oxigênio**, uso de antibiótico, estilo de vida, maioria do estudo, criança com idade, prevalência de asma, duração do aleitamento, volume de leite, aplicação da vacina, necessidade de ventilação, aumento da pressão, saúde da criança, **risco de infecção**, maioria da criança, uso em criança, **transtorno de ansiedade**, **modelo de regressão**, idade da criança, período de tempo, índice de massa, **ingestão de cálcio**, criança do sexo, **radiografia de tórax**

Trigrama SN

fator de risco, **aleitamento materno exclusivo**, profissional de saúde, **leite de vaca**, peso de nascimento, critério de inclusão, coleta de dado, serviço de saúde, **tipo de parto**, país em desenvolvimento, tempo de internação, polissacarídeo de soja, **faixa etária pediátrico**, trabalho de parto, **ventilação pulmonar mecânica**, **terapia intensivo neonatal**, casca de banana, uso de antibiótico, produção de leite, **saturação de oxigênio**, **intervalo de confiança**, velocidade de crescimento, qualidade de vida, terapia intensiva pediátrica, **ponto de corte**, **otite médio agudo**, ventilação não invasivo, nível de significância, consumo de medicamento, **dieta de exclusão**, sala de parto, uso de oxigênio, **lesão pulmonar agudo**, estilo de vida, equipe de saúde, acidente de transporte, **radiografia de tórax**, **grupo de risco**, farelo de trigo, análise de variância, termo de consentimento, **fórmula de soja**, critério de exclusão, densidade mineral óssea, **suplemento de cálcio**, **taxa de mortalidade**, uso de medicação, diferença estatística significativa, **doença de base**, selo de água, **relaxamento muscular inadequado**, uso de medicamento, **centro de referência**, **vacina contra influenza**, década de 70, **ansiedade de separação**, deficiência de vitamina, criança mais velha, estudo de corte, escape de ar, nível de linfócito, esquizofrenia com início, **curva de crescimento**, tempo de queixa, **risco de infecção**, **aspiração de mecônio**, uso de chupeta, **amostra de sangue**, primeiro 24 hora, **resposta inflamatório sistêmico**