Indexação e Recuperação Automática de Textos Médicos

Stefan Schulz[1], Percy Nohama[2], Emerson Paulo Borsato[2], Lúcio Jorge Dias Matias[2]
[1]Departamento de Informática Médica, Universidade de Freiburg (Alemanha)
[2]Programa de Pós-Graduação em Informática Aplicada, Pontifícia Univ. Católica do Paraná

Resumo - O acesso à informação médica deve ser eficiente, obtendo dados atualizados e confiáveis. Disso depende a qualidade dos serviços médicos prestados. Atualmente o modo de acesso a estas informações é realizado por meio de recuperação textual, tanto em bancos de informações *on-line* como *off-line*. Para se lidar adequadamente com esses gigantescos bancos de informações, é necessário que se utilize ferramentas de indexação e recuperação corretas e eficientes, que permitam tratar índices, palavras-chave e várias línguas. A metodologia usada atualmente, baseada em texto, é muito insatisfatória, pois não consegue tratar variáveis complexas próprias das expressões médicas. Esse artigo descreve o desenvolvimento de metodologia e ferramentas que abandonam os métodos tradicionais de recuperação e se baseia no uso de MORFEMAS médicos, como unidades atômicas para indexação e recuperação de informações

Palavras-chave: Indexação e recuperação de informação.

Abstract – The access to the clinical information must be efficient, obtaining trustfully data. This depends the quality of the medical care. Today, the way of this access is done by textual retrieval, in on-line databases and in off-line databases. To deal efficiently with these huge amounts of information, we must use correct and efficient indexing and retrieval tools. The methodology used today, based in text, is not satisfactory, because it cannot deal with complexes medical expressions. This article describes the development of a methodology and tools that abandon the traditional method of information retrieval and is based upon medical MOEPHEMES, as atomic units to the information indexing and retrieval.

Key-words: Information indexing and retrieval.

Introdução

A quase totalidade da informação médica produzida é expressa por meio da linguagem natural (texto livre), e o volume de informações disponíveis está crescendo a ponto de dificultar a seleção e a leitura do que é, de fato, útil e de interesse ou não. Assim, são necessárias ferramentas adequadas para que se possa recuperar eficientemente informações de fontes heterogêneas (Internet, redes locais, bancos de dados,...).

Motores de busca (como Google, AltaVista e outros) têm-se tornado utensílios indispensáveis na Internet e em redes locais, utilizando técnicas cada vez mais sofisticadas de recuperação de informação (information retrieval) [1,2]. Mesmo assim, suas limitações ainda são consideráveis: Dependendo da expressão submetida à busca, a quantidade de respostas pode ser extremamente elevada, embora a maioria dessas respostas sejam irrelevantes ou mesmo não pertinentes à informação desejada. Ou, pelo contrário, muitos documentos relevantes não são encontrados. Esse fenômeno explica-se pelo fato que motores de busca convencionais recuperam documentos na medida em que as expressões de busca ocorrem neles letra por letra. O *significado* das palavras e, em conseqüência, palavras semelhantes ou sinônimos não são levados em consideração.

Metodologia

Este artigo descreve uma metodologia que pretende superar essa restrição. A idéia central é extrair de um documento somente informações relevantes para a busca, as quais geralmente estão contidas nas raízes

das palavras (radicais) para construir o índice do documento, em vez de usar a "superfície" do texto – quer dizer, as palavras originais. Quando se efetua o tratamento do motor de busca pelas raízes das palavras e existe a tradução dessas raízes para outras línguas, torna-se possível fazer consultas utilizando termos em uma determinada língua e o motor de busca trazer respostas pertinentes em outra, dependendo do interesse do usuário.
Neste artigo, as informações estudadas são relacionadas ao contexto médico mas o método pode ser aplicado a quaisquer outros contextos. Isto exige uma análise detalhada da linguagem tal como é no contexto médico.

Mais do que a linguagem do dia-a-dia, a terminologia médica exibe características próprias que dificultam o uso eficiente dos mecanismos de busca:

- variação ortográfica: *diabetes mellitus, diabete mélito*
- derivação: *diabetes, diabético, diabéticas, antidiabéticos*
- composição: *hiperprebetalipoproteinemia,*
- sinonímia: *nephro*..., *renal*; *estômago*, *gastr*...;...
- abreviação: AVC, ECG, DPOC, ...
- nomes próprios: *diclofenaco*, *Viagra*, *Parkinson*, ...

Observa-se também, cada vez mais, a necessidade de lidar com documentos em línguas diferentes. Enquanto se usa a língua nativa para a documentação clínica, trabalhos científicos, geralmente, estão publicados em inglês. Diante desses fatos, os atuais métodos produzem recuperações incompletas e imprecisas, trazendo informações não pertinentes à pesquisa desejada.
Assim, surgem os seguintes desafios com relação à recuperação de textos médicos:

- a recuperação *intralingual* (na mesma língua nativa) e
- a recuperação *interlingual* (entre idiomas diferentes)

Dentro desse contexto, dois projetos encontram-se em desenvolvimento: o primeiro é um projeto de cooperação entre o Departamento de Informática Médica[a] da Universidade de Freiburg (Alemanha) e o Grupo de Tecnologia em Saúde do Programa de Pós-Graduação em Informática Aplicada (PPGIA)[b] da Pontifícia Universidade Católica do Paraná (PUCPR). Neste projeto, desenvolve-se o MORPHOSAURUS, uma metodologia que tem como objetivo aperfeiçoar a busca em coleções multilíngues de documentos médicos. O segundo projeto utiliza a mesma metodologia e está sendo desenvolvido pelo Grupo de Tecnologia em Saúde com o intuito de formar uma base de informações em doenças crônico-degenerativas para facilitar as pesquisas via World Wide Web (WWW). Este projeto também possui o apoio científico do Departamento de Informática Médica da Universidade de Freiburg.
A idéia principal é submeter todos os documentos a um processo de normalização morfossemântica antes de serem automaticamente indexados para, assim, melhorar o desempenho de motores de busca [3]. Este processo divide-se em:

- identificação de todos os componentes do texto aos quais se pode assinalar um significado atômico (não composto), com relevância para a busca de documentos. Essas chamadas *subwords* (sub-palavras) correspondem, na maioria dos casos, a morfemas. Como exemplo, o termo *miocardite* é composto pelas *subwords mio* (músculo) e *card* (coração), seguidos pelo sufixo *-ite* (inflamação);
- substituição de todas essas partículas assim identificadas por um identificador comum. Sinônimos (seja dentro da mesma língua ou entre línguas diferentes) recebem o mesmo identificador;
- indexação dos documentos assim preparados usando motores de busca (*crawlers*) convencionais.

As expressões de busca - digitados em texto livre pelo usuário - serão submetidas ao mesmo procedimento.

O sistema MORPHOSAURUS abrange as bases terminológicas e as rotinas de normalização de textos.

[a] http://www.imbi.uni-freiburg.de/medinf/mie_home.htm

[b] http://www.ppgia.pucpr.br/

As bases terminológicas são constituídas pelos seguintes componentes:

- um repositório de *subwords* classificados como radicais, prefixos e sufixos. A delimitação de subwords se efetua de tal modo que sirvam para a resolução de sinônimos, prevenindo, ao mesmo tempo a proliferação de ambigüidades no processo de segmentação do texto;
- um repositório de nomes próprios, tais como nomes de medicamentos ou nomes próprios, utilizados dentro de termos médicos, por exemplo, doença de *Alzheimer;*
- uma base de dados que mapeia acrônimos às *subwords* contidas na sua definição, e.g. ECG = *electr* + *card* + *graph*;
- um componente de *thesaurus* (dicionário de sinônimos) o qual agrupa *subwords* de significado idêntico assinalando-lhes um identificador comum;
- um mapeamento do repositório de *subwords* para o MeSH (Medical Subject Headings).

A terminologia interage com as seguintes rotinas de normalização de textos:

- segmentador: um programa que extrai *subwords* do texto para, depois, substituí-las no texto pelo seu identificador, usando o repositório de *subwords*.
- processador de acrônimos: acrônimos e abreviações são identificados e expandidos usando a base de acrônimos.

Resultados

Como resultado desse processo, obtém-se uma representação simbólica do conteúdo pela qual grande parte da variabilidade lingüística é eliminada. Essa representação, porém, mantém a estrutura original do documento e pode, por conseguinte, ser indexada por motores de busca convencionais.

Como exemplo, dois textos em português e inglês, de conteúdo idêntico. Depois de serem submetidos a um processo de segmentação, as *subwords* significativas são extraídas (em letras maiúsculas) e os acrônimos são identificados (entre aspas). Como exemplo, da palavra "*tireoideana*", o radical TIREOID é extraído; de "gestação", GESTAC; de "*pregnancy*" PREGNAN.

As *subwords* significativas são, então, substituídas pelo seu identificador (aqui usa-se um dos sinônimos em inglês como identificador); outras, como sufixos de flexão, artigos etc. são suprimidas. Como exemplo, todas as *subwords* {GESTAC, GESTANT, GRAVID, PRENH, PREGNAN} são substituídas pelo identificador "*pregnan*". Fragmentos que não contêm correspondentes no dicionário, são preservados: aqui "Hashimoto": HASH IM OTO.

Acrônimos são substituídos pela seqüência dos identificadores da sua definição, isto é, "TSH" por "*thyr stimul hormon*". A solução para acrônimos ambíguos consiste em substituí-los pela seqüência de todas as suas definições, isto é, "APA" por "*anti peroxyd enzyme antibody*" e "*america psychol associat*" .

Discussão e conclusões

Em sua versão atual, o sistema tem aproximadamente 15 mil *subwords*, abrangendo a terminologia clinica em inglês, alemão e português. A construção dos repositórios de nomes próprios e acrônimos ainda não foi abordada.

Um estudo controlado [4] (avaliando os parâmetros *precisão* e *recall*), baseado em uma amostra de textos alemães, mostrou que essa metodologia - mesmo usando uma implementação prototípica e uma versão ainda incompleta do vocabulário - teve um desempenho melhor que a abordagem convencional (indexação de palavras). Consequentemente, espera-se que com uma versão mais elaborada obtenha-se, também para o português, bons resultados, a fim de incorporá-la aos mais diversos ambientes de gerenciamento de conteúdo médico.

Referências:

[1] Hersh WR: *Information Retrieval – A Health Care Perpective*. New York: Springer, 1996.

[2] Rijsbergen CJ: *Information Retrieval*. London: Butterworth, 1979. www.dcs.gla.ac.uk/Keith/Preface.html
[3] Schulz S., Hahn U.: Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics*, 2000; 58-59: 87-99 http://www.elsevier.com/gej-ng/10/22/36/48/25/34/article.pdf
[4] Hahn U, Honeck M, Piotrowski M, Schulz S: Subword Segmentation - Leveling out Morphological Varieties for Medical Document Retrieval. In: Bakken S (Hrsg): *Visions of the Future and Lessons from the Past*. Proceedings of the 2001 AMIA Annual Symposium, Washington, November 3-7. Hanley & Belfus, 2001; 229-234

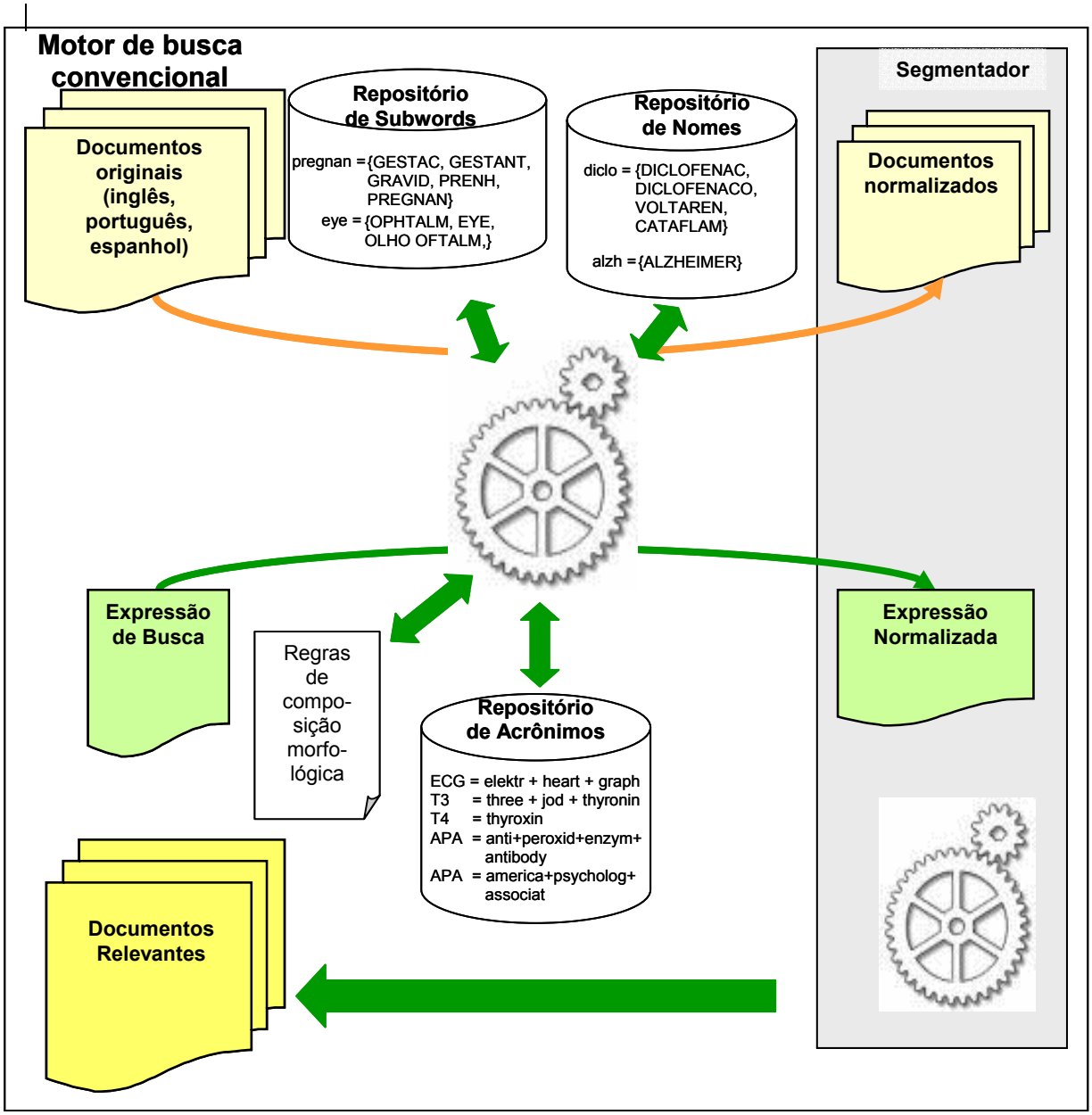

Figura 1. Arquitetura de MORPHOSAURUS.

Disfunção tireoideana perinatal

As doenças da tireóide acometem 10% das mulheres, mas a maioria das pacientes responde bem ao tratamento.

Durante a gestação, mudanças metabólicas podem ocultar a presença da patologia, com risco de dano fetal devido à conduta inapropriada. Os exames de *TSH*, *tiroxina livre* e *triiodotironina livre* são essenciais.

Geralmente, a presença de valores elevados de TSH sugere o diagnóstico de hipotireoidismo primário, enquanto níveis suprimidos de TSH sugerem hipertireoidismo. Este último costuma manifestar-se através de *bócio*, *oftalmopatia*, *fraqueza muscular*, *taquicardia* ou *perda de peso*.

Entre as mulheres, a etiologia mais freqüente de doença tireoideana é a auto-imunidade da tireóide (doença de Graves e tireoidite de Hashimoto); as mulheres com a doença apresentam um risco maior de disfunções tireoideanas no período pós-parto.

Usualmente, as mulheres com diagnóstico de doença de Graves durante a gestação já apresentavam sintomas de hipertireoidismo no período pré-concepção e, algumas vezes, as imunoglobulinas estimuladoras de tireóide podem elevar-se a ponto de induzir hipertireoidismo fetal. Já as mulheres com tireoidite de Hashimoto geralmente são eutireoideas mas, na presença de bócio difuso, podem ser hipotireoideas. O diagnóstico é confirmado através do achado de valores elevados de anticorpos de antiperoxidase. Outras formas de distúrbios tireoideanos são os nódulos benignos e malignos com hiperemese gravídica.

Perinatal Thyroid Dysfunction

Thyroid gland diseases affect 10% of women, but most patients respond well to treatment.

During pregnancy, metabolic changes can hide the presence of the disorder, with the risk of fetal damage due to inappropriate handling. Measurement of *TSH*, *free T4* and *T3* are indispensable.

Generally, high TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism. Typical manifestations of the latter are *goiter, ophtalmopathy, muscular weakness, tachycardy,* or *weight loss*

Among women the most common etiology of thyroid disease is thyroidal autoimmunity (Graves' disease and Hashimoto's thyroiditis. Women with this disease have a higher risk of thyroidal dysfunctions after birth.

Usually, women with evidence of Graves' disease during pregnancy already show symptoms of thyroid hypofunction before conception, and sometimes the thyroid stimulating immunoglobulins reach a level sufficient to induce fetal hyperthyroidism. However, women with Hashimoto's disease have a normal thyroid function. With a diffuse goiter they may exhibit a suppressed thyroidal function. The diagnosis is confirmed by high APA values. Other forms of thyroid disorders are benignant and malignant nodules with hyperemesis gravidarum.

DIS FUNCAO TIREOID e ana PERI NATAL

as DOENCA s da TIREOID e ACOMET em 10% das MULHER es MAS a MAIOR ia das PACIENT es RESPOND e BEM ao TRATAMENT o.

DURANTE a GESTAC ao MUDANCA s METABOL ic as PODEM OCULT ar a PRESENC a da PATOLOG ia COM RISC o de DANO FETAL DEVIDO a CONDUT a in APROPRIAD a. os EXAME s de *"TSH"*, *TIROXIN a LIVR e* e *TRI IODO TIRONIN a LIVR e* sao ESSENCI ais

GERAL mente a PRESENC a de VALOR es ELEVAD os de "TSH" SUGER e o DIAGNOST ic o de HIPO TIREOID ism o PRIMAR io ENQUANTO NIVEIS SUPRIM id os de "TSH" SUGER em HIPER TIREOID ism o. este ULTIM o COSTUM a MANIFEST ar se ATRAVES de *BOCIO, OFTALM o PATIA FRAQU eza MUSCUL ar TAQUI CARD ia* ou *PERD a de PESO*.

ENTRE as MULHER es a ETIOLOG ia mais FREQUENT e de DOENCA TIREOID e an a e a AUTO IMUN i dade da TIREOID e (DOENCA de GRAVE s e TIREOID ITE de HASH IM OTO); as MULHER es COM a DOENCA APRESENT am um RISC o MAIOR de DIS FUNC oes TIREOID e an as no PERIODO POS PARTO.

USUAL mente as MULHER es COM DIAGNOST ic o de DOENCA de GRAVE s DURANTE a GESTAC ao ja APRESENT ava m SINTOM as de HIPER TIREOID ism o no PERIODO PRE CONCEP cao e algum as VEZ es as IMUN o GLOBULIN as ESTIMUL ador as de TIREOID e PODEM ELEVAR se a PONTo de INDUZ ir HIPER TIREOID ism o FETAL. ja as MULHER es COM TIREOID ITE de HASH IM oto GERAL mente sao EU TIREOID e as mas na PRESENC a de BOCIO DIFUS o PODEM ser HIPO TIREOID e as. o DIAGNOST ic o e CONFIRM ad o ATRAVES do ACHA do de VALOR es ELEVAD os de ANTICORP os de ANTI PEROXID ASE. OUTR as FORM as de DISTURB io s TIREOID e an os sao os NODUL os BENIGNOS e MALIGN os COM HIPER EMES e GRAVID ic a.

PERI NATAL THYROID DYS FUNCTION

THYROID GLAND DISEAS es AFFECT 10% of WOMEN BUT MOST PATIENT s RESPOND WELL to TREATMENT

DURING PREGNAN cy METABOL ic CHANGE s CAN HIDE the PRESENCE of the DISORDER WITH the RISK of FETAL DAMAGE DUE to in APPROPRIAT e HANDL ing. MEASURE ment of *"TSH"*, *FREE "T4"* and *"T3"* are INDISPENSABLE

GENERAL ly HIGH "TSH" VALUE s SUGGEST the DIAGNOS is of PRIMAR y HYPO THYROID ism WHILE a SUPPRESS ed "TSH" LEVEL SUGGEST s HYPER THYROID ism. TYP ic al MANIFEST ation s of the LATTER are *GOITER, OPHTALM o PATHY, MUSCUL ar WEAK ness TACHY CARD y* or *WEIGHT LOSS*.

AMONG WOMEN the MOST COMMON ETIOLOG y of THYROID DISEAS e is THYROID al AUTO IMMUN ity (GRAVE s DISEAS e and HASH IM OTO s THYROID ITIS). WOMEN WITH this DISEAS e have a HIGHER RISK of THYROID al DYS FUNCTION s AFTER BIRTH.

USUAL ly WOMEN WITH EVIDENCE of GRAVE s DISEAS e DURING PREGNAN cy already SHOW SYMPTOM s of THYROID HYPO FUNCTION BEFORE CONCEPTION, and SOMETIMES the THYROID STIMUL at ing IMMUN o GLOBULIN s REACH a LEVEL SUFFIC i ent to INDUC e FETAL HYPER THYROID ism. HOWEVER, WOMEN WITH HASH IM OTO s DISEAS e have a NORMAL THYROID FUNCTION WITH a DIFFUS e GOITER they may EXHIBIT a SUPPRESS ed THYROID al FUNCTION. the DIAGNOS is is CONFIRM ed by HIGH "APA" VALUE s. OTHER FORM s of THYROID DISORDER s are BENIGN ant and MALIGN ant NODUL es WITH HYPER EMES is GRAVID ar um.

ill funct thyr about birth

patho thyr affect 10% femin but high patient respond good treatment.

during pregnan change metabol possibl hide present patho with risk damage fetus due behav suitabl. exam *thyr stimul hormon*, *thyroxin free three jod thyronin free* essential.

general present value high *thyr stimul hormon* suggest diagnos low thyr first during level suppress *thyr stimul hormon* suggest high thyr. last custom manifest by *goiter, eye patho weak muscle speed heart lose weigh*.

among femin etio high frequent patho thyr auto immun thyr (patho severe thyr inflamm hash im ear); femin with patho present risk high ill funct thyr period after birth.

usual femin with diagnos patho severe during pregnan present symptom high thyr period before conception vez immun globulin stimul thyr possibl high point induc high thyr fetus. femin with thyr inflamm hash im general norm thyr present goiter diffus possibl low thyr. diagnos by find value high antibody anti peroxyd enzyme. other form patho thyr nodul benign malign with high vomit pregnan.

about birth thyr ill funct

thyr gland patho affect 10% femin but high patient respond good treatment

during pregnan metabol change can hide present patho with risk fetus damage due suitabl manag. measure *thyr stimul hormon* , *free thyroxin three jod thyronin* essential

general high *thyr stimul hormon* value suggest diagnos first low thyr during suppress *thyr stimul hormon* level suggest high thyr. typ manifest last *goiter, eye patho, muscle weak speed heart weigh lose*.

among femin high common etio thyr patho thyr auto immun (grave patho hash im ear thyr inflamm). femin with patho high risk thyr ill funct after birth.

usual femin with evidence grave patho during pregnan show symptom thyr low funct before conception, sometimes thyr stimul immun globulin reach level sufficient induc fetus high thyr. however, femin with hash im ear patho norm thyr funct with diffus goiter show suppress thyr funct. diagnos confirm high anti peroxyd enzyme antibody america psychol associat value. other form thyr patho benign malign nodul with high vomit pregnan.

Figura 2. Exemplo de segmentação morfológica e normalização semântica.