3
Desenvolvimento de Dicionários Temáticos

Este capítulo trata da construção de dicionários temáticos para apoio à extração de entidades úteis em históricos policiais. Este capítulo encontra-se organizado de acordo com os tópicos seguintes: referencial teórico, método, algoritmos, sistemas, testes e avaliação do método para construção do dicionário especialista.

3.1
Referencial Teórico

Os históricos policiais podem ser classificados segundo a estrutura e organização dos dados armazenados e da forma como são coletados na origem.

Categorias de armazenamento relacionam-se com o nível de detalhamento de dados extraídos e das facilidades encontradas para representação de um Mapa de Inteligência. A classificação dos dados compreende três possíveis categorias:

- **Desestruturados** - são informações diretamente coletadas nos boletins de ocorrência em uma linguagem natural e característica do ambiente policial. Boletins desestruturados apresentam narrativas ricas dos fatos, objetos e instrumentos utilizados. Geralmente os documentos e arquivos de dados textuais permitem declarações referentes à dinâmica do crime, circunstâncias e envolvidos. Os dados coletados, devido à flexibilidade oferecida para os relatos das ocorrências, propiciam a introdução de erros, siglas, neologismos e abreviaturas desconhecidas.
- **Semi-estruturados** - são informações geralmente extraídas de outras bases digitais ou coletas parcialmente no registro da ocorrência criminal. Apresentam detalhes sintetizados da dinâmica do crime. A narrativa da ocorrência encontra-se restrita pela natureza do veículo de transcrição usado para captação do registro do delito criminal. Este tipo de estrutura oferece como maior vantagem a rápida identificação de atores, objetos, datas e localizações.

- **Estruturados** - Segundo Nardi e Wrigth (1998), dados estruturados são informações reconhecíveis pela gramática. Dados estruturados são pobres para investigação de dinâmicas criminais porque omitem detalhes coletados nas entrevistas e perdem informações valiosas das ações vinculadas às ocorrências policiais. Boletins policiais estruturados caracterizam-se por ofecerecem relativa facilidades para identificação da vítima, local, data e tipo da ação criminal desenvolvida, mas apresentam dificuldades para descoberta dos vínculos de autoria, descrição completa da dinâmica do crime, objetos e detalhes não contemplados no descritivo da ocorrência.

A eficiência da extração do conhecimento em bases de históricos policiais depende essencialmente da clara identificação dos atores nas fontes pesquisadas. A identificação está diretamente relacionada com a facilidade, segundo a qual as ferramentas de pesquisa encontram e identificam estes atores nos arquivos pesquisados e que função desempenham no cenário modelado. Quanto mais facilidades oferecer a fonte da pesquisa, mais direta e rapidamente serão encontrados os elementos desejados para modelagem de um Mapa de Inteligência criminal.

A desestruturação das informações, característica de boletins de ocorrências criminais, dificulta a identificação de atores mapeados, reduz a nitidez das fronteiras dos cenários delineados, interpondo-se como obstáculo para identificação de relacionamentos e confirmação de cumplicidades, consistindo em fator de retardo na apuração de responsabilidades (Xu e Chen, 2004). O fator tempo torna-se elemento fundamental para esclarecimento da autoria. Witten et al.(1999) citam ser possível a extração do conhecimentos úteis em um texto desestruturado, independentemente da compreensão de seu conteúdo pelo analista usando técnicas de mineração de dados e análise de textos por algoritmos.

Chen et al.(2004) descrevem que técnicas tradicionais de mineração de dados, como análise associativa de padrões (clusters), classificações de agrupamentos e previsões estavam restritas à bases estruturadas de dados. Presentemente técnicas de mineração de informações criminais foram expandidas para bases estruturadas e desestruturadas de informações (Chen et al., 2004). Dentre técnicas relacionadas com mineração de dados em bases desestruturadas

destaca-se o uso de dicionários de apoio, ferramenta auxiliar para identificação de conceitos úteis extraídas de coleções de históricos em formato textual.

3.1.1
Preparação dos Dicionários

Tesauro ou dicionário é uma estrutura de dados incorporada ao suporte de recursos destinados à mineração inteligente de dados em históricos digitais. Os dicionários são gerados a partir de informações classificadas, contidas em documentos pertencentes ao domínio da aplicação, que são convertidas em estruturas especializadas através de contínuo treinamento (Chen & Lynch, 1992). A pesquisa em dicionários representa um importante recurso em sistemas de recuperação de informações (Hull, 1996).

Dicionários compreendem informações selecionadas em documentos, bases de dados ou conceitos gerados manualmente por especialistas, que servem de orientação para algoritmos de extração como palavras-chave e rotinas de crítica e depuração.

Dicionários automáticos ou semi-automáticos podem ser gerados a partir de radicais processados por algoritmos que servirão de palavras-chave para extração de conhecimento. As estruturas iniciais processadas a partir dos radicais serão posteriormente especializadas através de treinamento executado por aprendizado de máquina (Porter, 1997). Dicionários manuais podem ser obtidos combinando palavras de domínio público (informações geográficas, profissões, siglas usuais, nomes comuns, títulos etc). As estruturas resultantes permitem que palavras-chave sejam extraídas de documentos textuais com pequena ou nenhuma interferência manual.

Vidal (2005) cita que o aprendizado de máquina utiliza algoritmos que permitem ao computador aprender com o ambiente ao qual está exposto, transformando regras que expressam o que há de importante nos dados. Chen & Lynch (1992) descrevem a geração automática de dicionários usando domínios residentes em bases bibliográficas.

Para construção dos dicionários são especificadas tabelas, regras e conceitos freqüentes, importantes para orientação dos algoritmos e processamento da extração. Os dicionários são geralmente temáticos, extensos e gradualmente

adquirem experiência para a área destinada através de treinamento para completar o ciclo de especialidade (Lippmann, 1987; Lee, 1998; Baluja et al.,1999). Lee (1998) cita um exemplo de palavras-chave para extração de conhecimentos na área anti-terrorismo. Neste modelo de aplicação, a busca procura entidades correlatas com assassinatos, seqüestros, reféns, bombas, explosivos, armas, treinamentos táticos e comunicações.

Dicionários servem de suporte às rotinas automáticas ou semi-automáticas de extração de entidades, padrões e regras relevantes de relacionamentos (Berry & Linoff, 1997).

A chave para extração de entidades é o reconhecimento de diferentes tipos de fragmentos textuais, denominados de *tokens ou entidades*. Um *Token* pode ser definido como um segmento de informação identificada (Lee,1998).

A construção de dicionários de suporte para o sistema extrator utiliza o método de processamento da "tokenização" de entidades em textos digitais, aplicando regras usadas para extração de conhecimentos e modelagem de redes, como palavras-chave e regras de pontuação, que são usadas como separadores léxicos. Baluja et al.(1999) consideraram as seguintes regras de pontuação para tratamentos de extração na língua inglesa: ponto, vírgula, travessão, exclamação, interrogação, ponto e vírgula, sinal mais, sinal menos, apóstrofo, parêntesis esquerdo, parêntesis direito. Baluja et al.(1999) utilizam um sistema de vetorização, contendo 29 sinalizadores associados à cada token. Cada elemento (sinal) associado recebe um valor binário, zero se ausente (falso) e um se presente (verdadeiro). Os sinalizadores representam a análise léxica da palavra, presença nos dicionários de suporte, pontuação, tipo de letra presente na palavra, nome próprio. Segundo Lee (1998), a extração constitui-se em uma meta linguagem de reconhecimento de padrões, representados por entidades textuais. A meta-linguagem seleciona qual o conteúdo e onde este deve ser procurado, tal como pessoas, nomes de empresas, localizações ou datas pesquisadas.

3.1.2
Métodos de Construção

Os dicionários para suporte à extração de conhecimento podem ser construídos usando três diferentes métodos:

- **Método manual**: os dicionários são alimentados a partir de dados pré-existentes, através da importação de informações residentes em outras bases de dados ou alimentados manualmente, através da inserção de dados gerados por profissionais com experiência nos domínios para os quais os dicionários são destinados. Por exemplo, dicionários de bases geográficas podem ser alimentados a partir de organizações como IBGE ou Fundação CIDE; dicionários para pesquisas de nomes de pessoas podem ser importados de cadastros policiais ou arquivos oficiais de Inteligência, como Fotocrim-Sinpol (2003).
- **Método Automático:** Quando o dicionário é gerado inteiramente a partir de extrações, via mineração de dados. O método automático de extração de dicionários geralmente envolve a combinação de algoritmos extratores de palavras-chave e treinamento contínuo da estrutura resultante para reciclagem e geração de especialidades. Chen & Lynch (1992) citam o uso freqüente de dicionários especializados em aplicações de mineração de dados.
- **Método Semi-automático:** o dicionário é gerado combinando-se extrações, via mineração e inserção manual, ou importação de informações a partir de outras bases de dados pré-existentes. Chen & Lynch (1992) citam a necessidade da cooperação de especialistas para complementação do domínio de conceitos usado para apoio à mineração como forma de garantir a praticidade dos resultados processados. Um dicionário que reflita de forma realista as palavras-chave para um determinado domínio temático pode demandar muitos homens /ano de exaustivo trabalho de construção manual.

3.1.3
Categorias e classificações de entidades

Um dicionário, cujo conteúdo é representado pelas palavras-chave extraídas relaciona-se fortemente com o domínio temático para o qual foi gerado.

Na Tabela 3.1, Lee (1998), Houston et al.(2000) e Klerks (2001), são apresentadas entidades relevantes representativas do conhecimento em distintos

domínios temáticos, baseados em aplicações anti-drogas, anti-terrorismo e área medicinal.

Tabela 3.1 – Exemplos de palavras chaves especializadas em domínios temáticos

Anti-Drogas	Anti-Terrorismo	Área Medicinal
Assassinatos Seqüestros Explosivos Tráfico de armas Demolições terroristas Planejamento Treinamento com armas e táticas terroristas Detenção de terroristas Ações de comandos anti-terrorismo	Processamento de material entorpecente Tráfico Compra e Venda de drogas Contrabando de entorpecentes Lavagem de dinheiro Planejamento Associações Detenção de traficantes Apreensão de drogas	DNA Enzima Mutação Autópsia Células Proteínas Anti-corpos Cobaias Tumor Morte Diagnóstico

Fonte: Lee, 1998, Houston et al., 2000 e Klerks, 2001

Os tipos de entidades selecionadas na construção de dicionários devem ser específicos para a finalidade da extração, bem como para o domínio temático onde serão utilizados.

Lee (1998) apresenta um exemplo de entidade, cujas propriedades dependem da natureza do modelo extrator. Para uma entidade <Indivíduo>, por exemplo, o sistema prevê: nome, apelido, gênero, estado civil, naturalidade, raça, data de nascimento, profissão, cor dos cabelos, cor dos olhos, altura e peso. Estas características podem, no entanto, variar com a finalidade do modelo extrator. Se esta mesma entidade estivesse inserida em um domínio médico, nas características previstas do indivíduo seria importante a inclusão de informações complementares do histórico médico e anamnese do paciente, tais como alcoolismo, diabetes, pressão arterial, cirurgias, tabagismo, doenças crônicas, etc. Para a área criminal seriam importantes informações do histórico criminal e atividades profissionais.

3.1.4
Restrições de linguagem e regras para extração de palavras-chave

O propósito da extração e idioma de referência impõe restrições e regras, que são específicos do modelo utilizado. Tais restrições referem-se a padrões

léxicos, sintáticos e regras morfológicas singulares, acrescidas de particularidades de linguagem encontradas com maior freqüência no domínio da pesquisa, tais como neologismos, gírias, siglas e abreviaturas. Segundo Lee (1998), as extrações de palavras-chave são específicas e restritas para o idioma para o qual dicionários e regras foram originalmente desenvolvidos (Inglês, Japonês, Português etc).

Baluja et al. (1999) citam, como exemplo, o uso de maiúsculas para representação de nomes próprios (exceto como inicial de uma frase) que pode variar de idioma para idioma: o chinês não tem maiúsculas; o espanhol apresenta nomes de localidades freqüentemente representados com minúsculas; o idioma alemão registra todos os nomes em letras maiúsculas, e finalmente, existem casos como van Gogh, combinando no mesmo nome letras iniciais minúscula e maiúscula. Baluja et al. (1999) ressaltam assim que nem todas as palavras em maiúsculas são nomes ou nem todos o nomes estão representados em maiúsculas.

Esforços para padronizações de procedimentos utilizados em extração de conhecimento têm sido desenvolvidos nas conferências MUC - Message Understanding Conferences (MUC-1 a MUC-7), desde a sua criação, em 1995. Estas conferências envolvem mensuração dos progressos no campo da extração de informações, tendo como foco principal de seus estudos:

- Reconhecimento de entidades nominais;
- Co-referência (freqüência de relacionamento entre entidades);
- Padronizações de procedimentos;
- Padronização de Cenários para extração de entidades.

Na MUC-7- *Message Understanding Conference* e *Second Multilingual Entity Task* (Chinchor, 1999), foi estabelecida uma classificação para **entidades**, que são conceitos extraídos por meio de sistemas de mineração de dados, representando nomes próprios, quantidades, pessoas, referências de locais, datas, horários, percentuais e valores monetários.

Baluja et al. (1999) apresentam **entidades** segmentados em três categorias, segundo padrões definidos na MUC:

- **Enamex** - para a identificação de nomes (pessoas, organizações etc);
- **Timex** - para a identificação de expressões temporais (datas e horários);
- **Numex** - para a identificação de expressões numéricas e quantificações (valores monetários, percentagens, etc).

3.1.5
Extração de palavras-chave em língua portuguesa

A geração e treinamento do dicionário especialista apresenta um viés no estudo da lingüística, que fornece as diretrizes básicas para o algoritmo de aprendizado do modelo de extração.

Quando um conjunto ordenado de fonemas apresenta um significado é designado como uma palavra. As palavras abrangem os nomes (substantivos, adjetivos e advérbios de modo) e os verbos (Monteiro, 2002).

Cognatos são palavras derivadas da mesma raiz, constituindo o elemento irredutível e comum a todas as palavras de uma mesma família (Monteiro, 2002), também denominada de família léxica (Laroca, 2005). O elemento é irredutível quando não pode mais ser segmentado.

Alguns exemplos de famílias que possuem a mesma raiz:

- Lua, enluarada, lunar
- Mar, maresia, marujo, marinheiro
- Crime, criminoso, criminologia, criminal
- Amor, amar, amigável, amigo, amizade, desamor

Monteiro (2002) cita que a estrutura interna das palavras é constituída de elementos associados que representam os elementos mínimos das emissões lingüísticas que contém um significado individual.

A parte fundamental ou núcleo significativo da palavra é o elemento que guarda o conteúdo significativo da série derivada, em sua forma mais primitiva. Monteiro (2002) cita como exemplo a palavra belíssima derivada de [bel] acrescido do sufixo [íssima]. Outros sufixos e prefixos agregados produziriam vocábulos como beleza, embelezar ou embelezamento.

O núcleo semântico da palavra é a parte comum a um grupo de palavras aparentadas pelo vínculo da significação. Assim o núcleo [caval] aparece em cavalo, cavalar, cavalaria, cavalete etc. Palavras derivadas podem ser criadas a partir da forma primitiva, como [caval] + [eiro] ou [aria]. Os sufixos [aria] ou [eiro] são derivacionais que permitiram a criação de palavras a partir da forma primitiva da palavra *cavalo*. A forma primitiva é o vocábulo de onde se originam

outros vocábulos através do processo de derivação. Os vocábulos derivados se denominam formas secundárias (Monteiro, 2002).

Uma palavra quando constituída de sufixos e/ou prefixos, é denominada de radical, como por exemplo em ***livraria***, com radical ***livr*** + ***aria*** (Laroca, 2005). Um radical pode então ser definido como uma estrutura constituída de raiz e afixos (prefixos e sufixos) (Laroca, 2005).

Monteiro (2002) cita que o radical em sua forma primitiva é a raiz, o elemento mínimo de uma família de palavras e elemento irredutível e comum desta família de palavras. A raiz é o elemento de onde parte a primeira operação morfológica, assim a raiz passa a ser diferente do radical. Os radicais podem ter um ou mais afixos derivacionais Desta forma uma mesma palavra pode ter diversos radicais. A palavra marinheiro pode oferecer três graus no radical:

1. mar
2. marinh
3. marinheir

Monteiro (2002) cita que o significado é essencial no conceito de raiz, portadora da carga semântica da palavra. Os sufixos especializam ou particularizam o significado genérico da raiz (parte mais reduzida da palavra) em uma série de derivados. O radical de grau mais elevado inclui todas as palavras derivadas. (Monteiro, 2002).

A contínua interação ***texto → radicais → dicionário → texto*** propicia uma visão estatística de palavras-chave mais comuns, presentes no domínio de informações pesquisado.

3.1.6
Extração de Radicais (*Stemming*)

Stemming é o processo de concentração de diversas formas da palavra em uma representação comum - o radical (Orengo & Huyck, 2001). Diversos algoritmos *stemming* têm sido gerados como ferramenta para extração de conhecimentos de textos, com propósito de reduzir palavras às suas formas radicais através de diferentes métodos, que compreendem remoção de sufixos,

supressão estrita de caracteres, segmentação de palavras, bigramas[1] e morfologia lingüística (Hull, 1996).

O algoritmo *Stemming* processa a remoção de sufixos e terminais inflexivos de palavras em Inglês como plurais, gerúndios e terminais específicos como "ator", "ate" e outros (Porter, 1997). A redução de palavras à forma de radicais permite o uso de vocábulos primitivos anteriores às variações, como plurais e inflexões verbais. Por exemplo, as formas *learned* e *learning* são reduzidas para *Learn* (Porter 1997).

Um dos problemas evidenciados por Hull (1996) em seu estudo de avaliação de algoritmos *Stemming* aponta para perda de significado de algumas palavras submetidas ao processo de radicalização, onde diferentes interpretações são concentradas no radical mais simplificado (reduzido). Como regra geral podemos afirmar que quanto mais reduzido estiver o radical, mais genérica será a extração. Por exemplo, através do algoritmo de Porter, *geral* (*general*), *generoso* (*generous*), *geração* (*generation*) e *genérico* (*generic*) são reduzidos para um significado único. Dentre algumas técnicas utilizadas, Hull (1996) cita que radicais produzidos por remoção de sufixos freqüentemente não constituem palavras, não servindo a outro propósito senão recuperação de informações em textos. Por outro lado, técnicas para utilização de redutores (palavras truncadas) selecionadas para expansão (acréscimos de afixos) apresentam grande performance quando aplicam radicais, em substituição a palavras reais (Hull, 1996). Um algoritmo típico para redução de palavras analisa e reduz a palavra para a sua forma morfologicamente inflexional e derivacional, conforme encontra-se apresentada em um dicionário léxico comum. Vide o exemplo apresentado na Tabela 3.2:

Tabela 3.2 – Regras de reduções

Palavra	**Reduzido para**	**Exemplo**
Substantivos	singular	Crianças → criança
Verbos	infinitivo	Compreendido → Compreender
Adjetivos	Forma positiva	Melhor → bom
Pronome	nominativo	A quem - quem

Fonte: Hull, 1996

[1] Seqüência de duas letras consecutivas ou dois números consecutivos.

O algoritmo *Stemming* pode transformar o radical em um tamanho muito reduzido. Por esta razão algoritmos *Stemming* (Porter, 1997; Lovins, 1998; Orengo et Huyck, 2001) testam os resultados da redução, não permitindo que os radicais ultrapassem um tamanho mínimo, geralmente fixado em três caracteres, dependendo do sufixo associado à regra aplicada ao ciclo de redução.

Algoritmos estudados (Porter, 1997; Lovins, 1998; Orengo & Huyck, 2001) trabalham com regras aplicadas ao tamanho mínimo do tamanho da palavra a ser reduzida.

Stem é o conjunto de caracteres resultante de um procedimento de *stemming.* O conjunto resultante não é necessariamente igual à raiz lingüística.

Chaves (2003) cita dois erros típicos que costumam ocorrer durante o processo de *stemming*:

- *Overstemming*: quando a cadeia de caracteres removida não é um sufixo, mas parte do stem.
- *Understemming* ocorre quando um sufixo não é removido completamente.

3.2
Algoritmos Pesquisados

Este tópico trata dos algoritmos pesquisados para construção de dicionários de apoio à extração de entidades de utilidade para modelagem de associações.

3.2.1
Algoritmo PORTER

Porter (2008) desenvolveu uma técnica baseada em regiões pré-identificadas nas palavras, sobre as quais o seu algoritmo está apoiado. Este algoritmo extrai radicais utilizando técnicas para tratamento de sufixos derivacionais, referenciando regiões. Porter (2008) denominou estas regiões de **R1 , R2 e RV.**

A palavra é segmentada em duas áreas: a região *anterior* ou pré-região **R1** e região *posterior* a R1, que compreende a região **R2**.

R1 compreende a região posterior à primeira consoante que segue uma vogal. Esta região será nula (Ø) ao final da palavra se esta condição não for atendida.

R2 é a região posterior à primeira consoante que segue uma vogal em **R1**. Esta região será nula (Ø) se esta condição não for atendida.

RV é uma região, cuja definição depende do idioma de referência. No caso dos idiomas Espanhol e Português, teremos:

- Se a segunda letra for uma consoante, **RV** compreenderá a região posterior à vogal seguinte.
- Se as primeiras duas letras são vogais, **RV** será a região seguinte à próxima consoante.
- No caso *consoante-vogal*, **RV** será a região posterior à terceira letra.

Os exemplos apresentados na Figura 3.1 evidenciam as regiões **R1** , **R2** e **RV**, representadas para palavras em Inglês:

Exemplo	Extrações de regiões
b e a u t i f u l	[R1] - **iful** [R2] - **ul** [RV] - **utiful**
A n i m a d v e r s i o n	[R1] - **imadversion** [R2] - **adversion** [RV] - **madversion**
B e a u t y *(R1 nula) / (R2 nula)*	[R1] - **y** [R2] - **Ø** [RV] - **uty**
b e a u *(R2 nula)*	[R1] - **Ø** [R2] - **Ø** [RV] - **u**

Figura 3.1 – Exemplos de regiões usadas na extração de sufixos R1, R2, RV (Porter, 2008)

O algoritmo Stemming Porter é utilizado para redução de radicais em diversos idiomas (Porter, 2006). Este algoritmo foi originalmente escrito em linguagem BCPL, linguagem extinta implementada no MIT em 1967. A versão do algoritmo Porter para o idioma Inglês (Porter, 2008) é executado em oito passos que são apresentados no Apêndice A.

Porter (2008) gerou ainda a meta-linguagem **SnowBall,** stemming algorithm language, para extração de palavras, com a qual algoritmos de extração podem ser traduzidos em diversos idiomas.

Snowball trabalha basicamente com as regiões R1, R2 e RV das palavras, associando sufixos e regras de substituição ou de transformação destas regiões por novos sufixos reduzidos ou supressão dos sufixos originalmente residentes. O resultado da especificação na meta-linguagem é a sua tradução em um programa em linguagem Java ou ANSI C.

Foram desenvolvidas versões usando a meta linguagem SnowBall para aplicações em outros idiomas como Espanhol, Português, Italiano, Romeno, Alemão, Dinamarquês, Sueco, Norueguês, Russo, Filandês, Húngaro e Turco (Porter, 2007).

3.2.2
Algoritmo ORENGO e HUYCK

Orengo et Huyck (2001) desenvolveram um algoritmo *stemmer* para língua portuguesa, definido parâmetros em uma meta-linguagem denominada **RSLP** *Stemmer* (*Removedor de Sufixos da Lingua Portuguesa*). Este algoritmo é executado em oito passos e 198 regras de redução e exceções.

Cada regra do algoritmo Orengo et Huyck contém quatro tipos de parâmetros:

- **Sufixo residente a ser removido** - especificação do sufixo que, se encontrado na palavra, será removido ou substituído.
- **Tamanho mínimo do radical residual** - exigência do número mínimo de caracteres residuais da palavra, se aplicada a regra de remoção.
- **Sufixo de substituição** - sufixo substituto ou nenhum, após aplicação da regra de remoção ou substituição.
- **Exceções** - Grupo associado de palavras que constituem exceções da regra de remoção ou substituição (as exceções não são obrigatórias nas regras de remoção)

O uso do algoritmo *stemming* em português é mais problemático que o algoritmo *stemming* em inglês devido à complexidade morfológica da língua portuguesa. Dentre as dificuldades encontradas, Orengo et Huyck (2001) citam:

- Exceções léxicas: o sufixo [ão], regularmente usado com aumentativo [casa casarão, Carro carrão], pode ser encontrado em palavras sem significado aumentativo [cão, televisão etc].
- Homônimos - palavras iguais na forma escrita e de diferentes significado (Aurélio, 1977) [cedo → advérbio, cedo → verbo].
- Homógrofos - palavras iguais na forma escrita, de diferentes significado e sons diferentes (Aurélio, 1977) [concerto → consonância de instrumentos, concerto → ato de concertar].
- Verbos irregulares
- Troca da raiz morfológica - causados normalmente pela inflexão da palavra [emitir → emissão: emit → emis]
- Nomes Próprios - devido a conflitos naturais e das múltiplas formas de tratamento dos nomes próprios, estes não devem ser reduzidos à radicais [Pereira → sobrenome , Pereira → árvore frutífera].

Uma ilustração do algoritmo Stemming para a língua portuguesa (Orengo & Huyck, 2001) pode ser encontrada no Apêndice B.

3.2.3
Algoritmo KEA

Witten et al. (1999 (b)) e Jones & Paynter (2001) propuseram o algoritmo KEA (*Key Phrase Extraction Algorithm*) para extração automática de conceitos usando palavras-chave de textos da língua inglesa. O KEA identifica palavras-chave usando análise léxica de frases, aprendizado de máquina para treinamento e extração automática de novas palavras-chave. São realizados três passos para execução do algoritmo KEA: limpeza do texto de entrada, identificação de frases candidatas e radicalização das frases selecionadas.

A principal característica do algoritmo KEA é a utilização de palavras-chave definidas por autores de pesquisas, que são posteriormente utilizadas para inicialização do algoritmo e extração de novos conceitos. A especialização de

dicionários pode ser obtida através de treinamento utilizando aprendizado de máquina (Baluja et al., 1999), técnicas de redes neurais (Chau et al., 2002; Vidal, 2005, Oatley, 2003), construção manual ou desenvolvimento automático, com auxílio de algoritmos (Feldman et al., 1998). Para o sistema de aprendizado de máquina (Baluja et al., 1999), o modelo de treinamento extrai cada frase do documento usado no treinamento, descartando frases duplicadas ou aquelas que iniciam ou terminam com *Stop-words*, palavras que ocorrem apenas uma vez ou ainda, frases com significado distante das palavras- chave conhecidas (Jones & Paynter, 2001).

Em cada documento processado na fase de treinamento, o texto é pesquisado e palavra-chave são identificadas. O algoritmo então gera um modelo que prediz a classe de uma determinada palavra (Lacerda & Gomensoro, 2004). Uma vez treinado, o modelo de reconhecimento de conceitos pode ser usado para extrair conhecimentos de novos documentos.

3.2.4
Algoritmo LOVINS

Publicado originalmente em 1968 por Julie Lovins, o algoritmo é executado em dois passos. No primeiro passo uma longa lista de sufixos maiores é testada para remoção. No segundo passo, 35 regras são aplicadas para transformação do terminal da palavra. O segundo passo não depende do primeiro passo ser executado para a sua consecução.

Para cada sufixo o algoritmo associa uma condição para remoção. A Tabela 3.3 apresenta um exemplo das regras de remoção e seus respectivos códigos.

Tabela 3.3 – Lista parcial de regras aplicadas ao algoritmo Lovins

Código da regra	Regra
Regra A	Nenhuma restrição aplicada ao radical
Regra B	Tamanho mínimo do radical = 3
Regra C	Tamanho mínimo do radical = 4
Regra E	Não remova este sufixo após a letra "e"
Regra H	Remova apenas após a letra "t"
Regra H"	Não remova este sufixo após as letras "*a*","*c*" *e* "*m*

Fonte: Lovins, 1968

Existem 35 regras aplicáveis no passo 2 do algoritmo Lovins, denominados de regras de transformação. As regras aplicam-se à terminação do radical, após a finalização do Passo 1 do algoritmo.

A Figura 3.2 apresenta uma ilustração do algoritmo Lovins para o idioma Inglês, (Lovins, 1998).

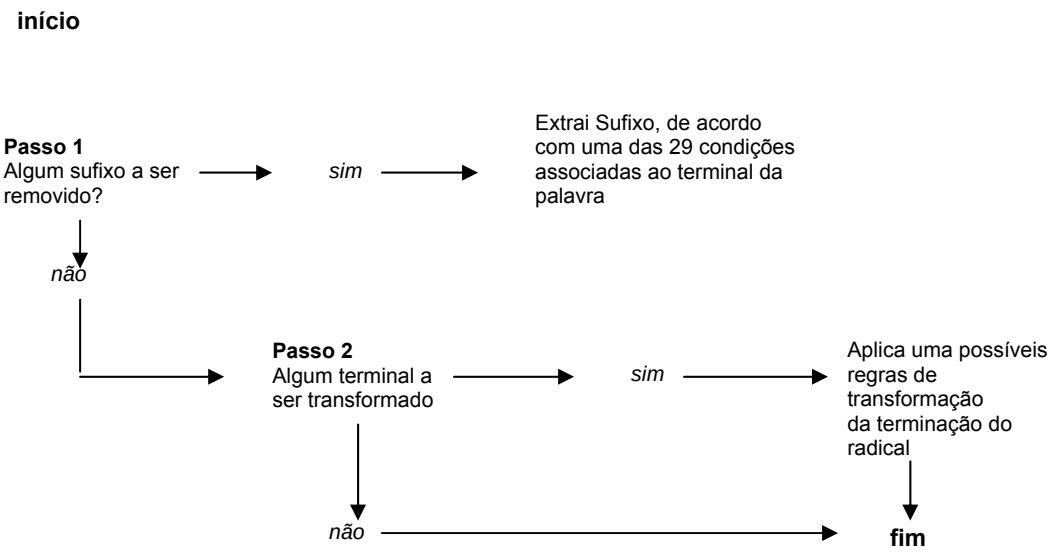


Figura 3.2 – Algoritmo Lovins para o idioma inglês, (Lovins, 1998).

3.2.5
Limitações dos algoritmos

O algoritmo Lovins (1998) foi influenciado pelo vocabulário técnico de sua autora Julie Lovins, o que determinou limitações na operação do algoritmo. Alguns sufixos foram suprimidos na construção do algoritmo Lovins, restringindo parcialmente o seu pleno uso para todos os vocábulos do idioma Inglês. Por

exemplo, o sufixo ***ements*** e a sua forma singular ***ents*** estão ausentes das lista de sufixos cadastrados no algoritmo Lovins (1998). Outra restrição do algoritmo é o tratamento de palavras pequenas, onde o algoritmo apresenta a tendência de ser destrutivo.

Hull (1996) cita que os algoritmos Porter & Lovins situam-se como algoritmos mais populares para recuperação de informações, sendo ambos baseados em remoção de sufixos.

O Algoritmo Lovins (1998), apesar de citado como mais extenso que o algoritmo de Porter (1997), pois apresenta uma lista ampliada de terminações, é mais rápido - necessitando apenas dois passos para remoção dos sufixos, contra oito passos exigidos pelo algoritmo de Porter (1997).

Hull (1996) cita como principal diferença entre os algoritmos Porter e Lovins o número de sufixos associados e regras utilizadas na extração de radicais. A Tabela 3.4 apresenta a síntese das diferenças que caracterizam os dois algoritmos:

Tabela 3.4 – Comparação entre os algoritmos Porter (1997) & Lovins (1998)

Algoritmo	**Descrição**	**Etapas**
Porter (1997)	Utiliza um algoritmo interativo , com menor número de sufixos e regras aplicadas ao contexto da extração	Executa a extração em oito passos
Lovins(1998)	Utiliza uma série de comparações contra uma lista maior de sufixos.	Executa a extração em dois passos

Fonte: Porter (1997) & Lovins (1998)

Krovetz (1997) cita que os métodos utilizados por Lovins (1998) e Porter (1997) podem gerar problemas de radicais homônimos (*conflação*).

Conflação é o processo de fusão ou agrupamento de termos para igualar variações morfológicas. A conflação pode ser realizada pelo processo manual, usando algum tipo de expressão regular, ou automática, via programas para extração de radicais (Chaves, 2003).

Em algoritmos voltados para extração de radicais desprovidos de sistemas de apoio léxico, o significado estrito do radical é ignorado, podendo acarretar erros de identificação e agrupamento de raízes. Krovetz (1997) cita que, devido ao

problema de homônimos, o algoritmo Porter comete erros agrupando "falsos amigos" morfológicos, como author / authority ou police / policy, entretanto deixa de agrupar palavras como *recognize* e *recognition.*

Hull (1996) destaca um exemplo de *conflação* produzido pelo algoritmo Porter: *general, generous, generation e generic,* "falsos amigos" agrupados sob a mesma raiz "*gener*".

Chaves (2003) desenvolveu um experimento envolvendo o algoritmo Orengo e Huyck, utilizando 500 palavras em Português com objetivo de aferir a sua precisão. O processo de análise foi executado manualmente, permitindo verificar o radical correto de cada palavra e posterior verificação dos resultados com a execução do algoritmo e aferição dos resultados. A Tabela 3.5 apresenta os resultados computados para a análise de aferição do algoritmo Orengo e Huyck desenvolvido por Chaves (2003).

Tabela 3.5 – Resultados computados no experimento de aferição do algoritmo

Categoria Gramatical	**Eficiência**	**Quantidade**	**Acertos**	**Over Stemming** (parte do radical é removido com o sufixo)	**Under Stemming** (sufixo não é completamente removido)
Substantivo	79,4%	263	209	44	8
Verbo	77,6%	134	104	13	6
Adjetivo	77,9%	59	46	9	4
Pronome	73,3%	15	11	3	0
Contração de preposição	57,1%	14	8	6	0
Advérbio	80,0%	15	12	3	0
Total (Abs.)	77,8%	500	389	77	18

Fonte: Chaves (2003)

Principais conclusões do experimento desenvolvido por Chaves (2003):

- O algoritmo Orengo e Huyck apresentou somente 18 erros atribuídos à understemming;
- Para a categoria gramatical substantivo, o algoritmo Orengo e Huyck cometeu maior volume de erros de overstemming ;

- O percentual total de acertos do algoritmo Orengo e Huyck foi de 78%;
- O algoritmo Orengo e Huyck processou 57,2% das palavras corretamente;
- O algoritmo Orengo e Huyck apresentou a melhor precisão para a categoria substantivo, processando 70% das palavras corretamente.

Considerando resultados observados em experimentos anteriores que utilizaram o algoritmo Orengo e Huyck, acrescidos do resultado verificado em seu experimento, Chaves (2003) concluiu pela dificuldade de se obter uma precisão na extração próxima aos 100% esperados como eficiência do algoritmo.

3.2.6
Uso de Dicionários para descoberta de Interseções

Feldman & Dagan (1996) citam a utilização de dicionários visando a descoberta de interseções de entidades em documentos, identificadas através de palavras-chave extraídas de coleções de documentos. A extração de palavras-chave para construção de dicionários permite a identificação de interseções e possíveis padrões existentes, dentre outras:

- Resumo estatístico de padrões encontrados, com base nas freqüências computadas;
- Identificação das categorias de padrões, segundo distribuições de freqüência;
- Identificação de sub-conjuntos de documentos, associados às categorias das distribuições de freqüências;
- Comparações entre sub-conjuntos de documentos, com base nas distribuições de freqüências processadas;
- Análise de tendências: comparações das distribuições de freqüências sob diferentes datas, gerando resultados sob forma de gráficos;
- Descoberta de associações entre classes de documentos;
- Referência cruzada entre freqüências e documentos que contribuíram para o seu somatório.

A busca pelo conhecimento permite ao usuário acessar documentos e reconhecer padrões baseados nas distribuições de freqüências de palavras-chave.

Feldman & Dagan (1996) citam uma pesquisa por padrões, extraídos de históricos médicos, com intuito de identificar um perfil clínico de pacientes, associado à histórico de internações, tendo como fundamentação um estudo de freqüência de co-ocorrências extraído de coleções de fichas médicas.

3.3
Descrição do Método para Construção de Dicionários

A primeira fase do método proposto nesta Tese trata da construção de dicionários que darão suporte à modelagem de Mapas de Inteligência. Além de um dicionário especialista obtido através de algoritmo, o método utiliza um conjunto de dicionários de suporte à extração de entidades nominais, tais como logradouros, nomes próprios e um dicionário de *Stop-words* representando conceitos descartáveis.

O pré-processamento do dicionário especialista está baseado no princípio de extração de radicais oriundos de conjuntos de treinamento. As listas extraídas são então utilizadas como palavras-chave, de forma interativa, obtendo-se palavras derivadas. A estrutura resultante é depurada, obtendo-se contribuições da base anterior existente e extraindo contribuições de palavras negativas ou Stop-words, que são palavras com pouco significado no texto analisado (Antiqueira et al., 2003).

O dicionário especialista é a principal estrutura de suporte à extração de entidades, sendo integrado por palavras freqüentemente encontradas em documentos manipulados durante as investigações criminais. Este dicionário é construído a partir de conjunto de documentos selecionados para treinamento, acrescido de contribuições complementares geradas por policiais, constituídas por um conjunto de palavras comuns encontradas em históricos criminais. A base de construção do dicionário especialista são radicais extraídos através de algoritmo, que fornecem uma semente lingüística para análise de textos e captação da linguagem comum utilizada nos documentos policiais.

O dicionário especialista será considerado como treinado quando for suficientemente preparado para reconhecer e extrair palavras-chave de quaisquer conjuntos de documentos temáticos, em uma proporção considerada como

eficiente, não recusando ou excluindo nesta extração as palavras-chave principais identificadas em uma investigação criminal.

A construção do dicionário especialista é desenvolvida em duas etapas. Em cada etapa são executadas extrações e treinamento dos resultados, em um processo cujo objetivo é especializar gradualmente a estrutura de informações resultante até uma proporção considerada como eficiente para extração de palavras-chave temáticas.

A construção do dicionário especialista segue o método seguinte, desenvolvido em dois passos:

- **Passo 1** - Construção de uma estrutura de radicais a partir de históricos policiais
- **Passo 2** - extração de palavras-chave que integrarão o dicionário especialista

Método de Construção do dicionário especialista - Passo I

Neste tópico é descrito um método para construção de uma estrutura de radicais a partir de palavras inseridas em uma coleção de documentos textuais selecionados. Os radicais serão posteriormente usados como base para extração de palavras-chave usadas na construção de um dicionário temático especializado

A estrutura de radicais é baseada em um algoritmo que elimina a região **RV** das palavras selecionadas.

Denominamos este algoritmo de **Anaphora RV** (trata a região RV da palavra).

O método permite a especialização de uma estrutura de radicais, aproximando tanto quanto possível o resultado extraído do vocabulário encontrado em documentos policiais. A estrutura de radicais resultante servirá como suporte ao reconhecimento e extração de palavras-chave temáticas em históricos policiais que serão utilizadas para construção de um dicionário especialista.

A especialização da estrutura de radicais extraídos tem como objetivos:

- Obter um volume máximo de radicais familiares com vocabulário presente em documentos policiais
- Apoiar de forma eficiente algoritmos voltados para extração de palavras-chave para construção de um dicionário especialista.

Método de Construção do dicionário especialista - Passo II

Neste passo são extraídas palavras-chave para geração de um dicionário temático especializado. A extração de palavras-chave apóia-se na estrutura de radicais selecionados no Passo I que são utilizados como sementes da extração. O método utiliza históricos policiais típicos, fonte da extração das palavras-chave derivadas dos radicais.

Selecionamos para desenvolvimento do dicionário especialista um algoritmo que pesquisa e extrai palavras-chave de documentos textuais livres, com base na estrutura de radicais treinados. Denominamos este algoritmo de **Anaphora PCh** (tratamento de palavras-chave) e uma lista de palavras indesejáveis à extração denominada stop-words..

A especialização do dicionário de palavras-chave tem como objetivos:

- Obter um volume máximo de palavras identificados em documentos policiais;
- Apoiar de forma eficiente algoritmos voltados para extração de entidades destinadas a modelagem de redes semânticas.

A Figura 3.3 apresenta o modelo de extração de radicais e construção do dicionário especialista de palavras-chave.

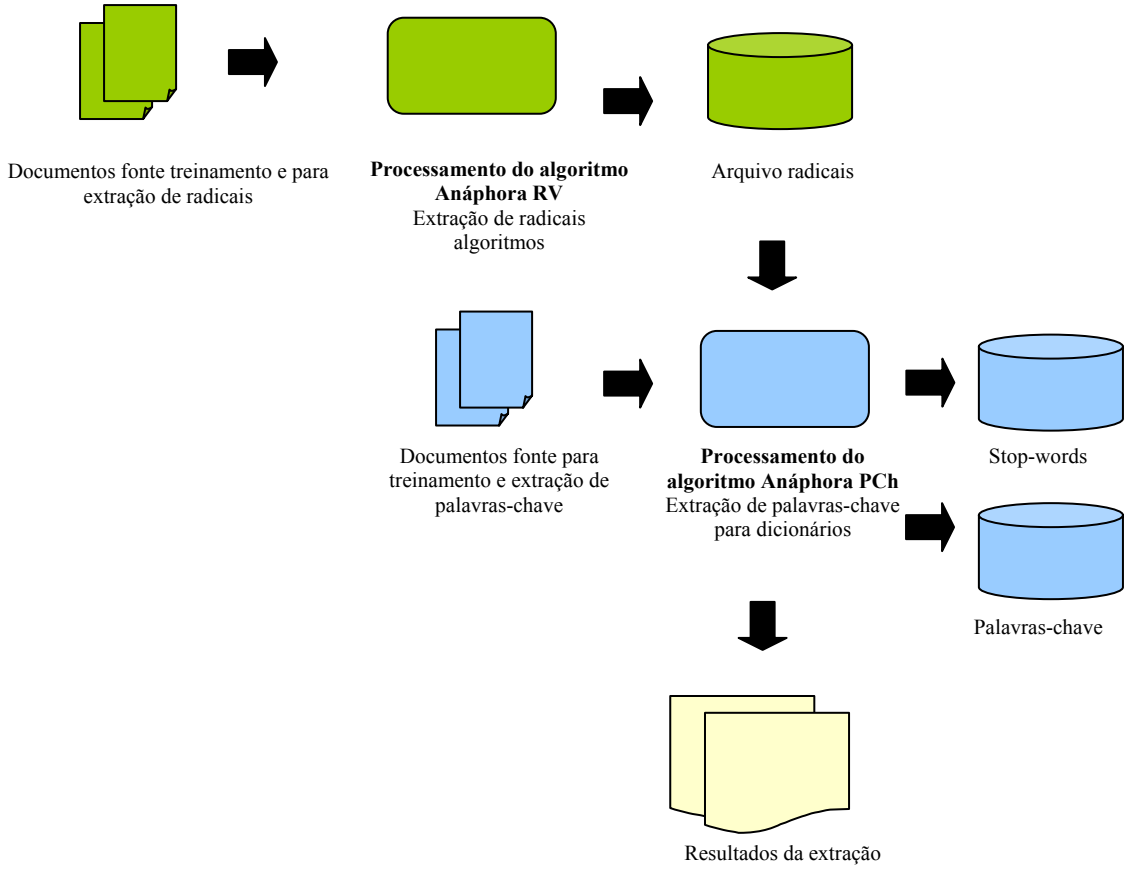


Figura 3.3 – Fluxo do processamento para extração de palavras-chave do dicionário especialista

3.4
Algoritmos utilizados para construção do dicionário especialista

Dois algoritmos foram usados para construção do dicionário especialista. O primeiro destes é um algoritmo do tipo Stemmer, que trata a extração de radicais de conjuntos de documentos, denominado Anaphora RV. O segundo, denominado Anaphora PCh, é responsável pelo tratamento e extração de palavras chave de conjuntos de documentos e utiliza a estrutura produzida no algoritmo Anaphora Rv.

3.4.1
Algoritmo Anaphora RV para Extração de Radicais

O algoritmo **Anaphora Rv** é executado em quatro passos:

- **Passo-1** - Limpeza do Texto e Geração de Tokens[2]

  O primeiro passo seleciona Tokens do texto de treinamento, extraindo todos sinais, pontuação e não-palavras.

- **Passo 2 -** Eliminação dos Tokens Inúteis (Stop Words)

  Os Tokens residuais são confrontados contra uma lista de Stop-words, eliminando-se aqueles, porventura, presentes nesta lista.

- **Passo 3 -** Exclusão das Maiúsculas

  Uma análise sintática é executada para identificação de palavras maiúsculas simples e maiúsculas combinadas (conectadas pelas preposições *da, de, do*). Estas palavras são eliminadas da lista de Tokens selecionados formando uma lista de candidatos potenciais para o dicionário de nomes próprios ou da lista de Stop-words.

- **Passo 4 -** Eliminação da Região RV dos Tokens residuais

  Extrai a região **RV**[3] dos Tokens residuais. O resultado da extração é o Radical-Chave.

- **Fim -** Reciclagem

O arquivo de radicais-chave obtido como produto do algoritmo **Anaphora Rv** retorna para inclusões de novos radicais, como parte do treinamento e especialização de conteúdo. Este ciclo se encerra quando o seu conteúdo é considerado eficiente para extrações de palavras-chave com mínimas perdas ou erros na extração.

O fluxo apresentado na Figura 3.4 apresenta os passos para desenvolvimento do algoritmo **Anaphora RV**.

[2] Token - unidade mínima de informação reconhecida no texto por mecanismos extratores

[3] Região RV - mediante regras específicas, a região RV divide a palavra em duas partes. Extraindo-se a região RV, que compreende a parte final da palavra, geralmente o resíduo inicial compreende entre dois a quatro caracteres iniciais.

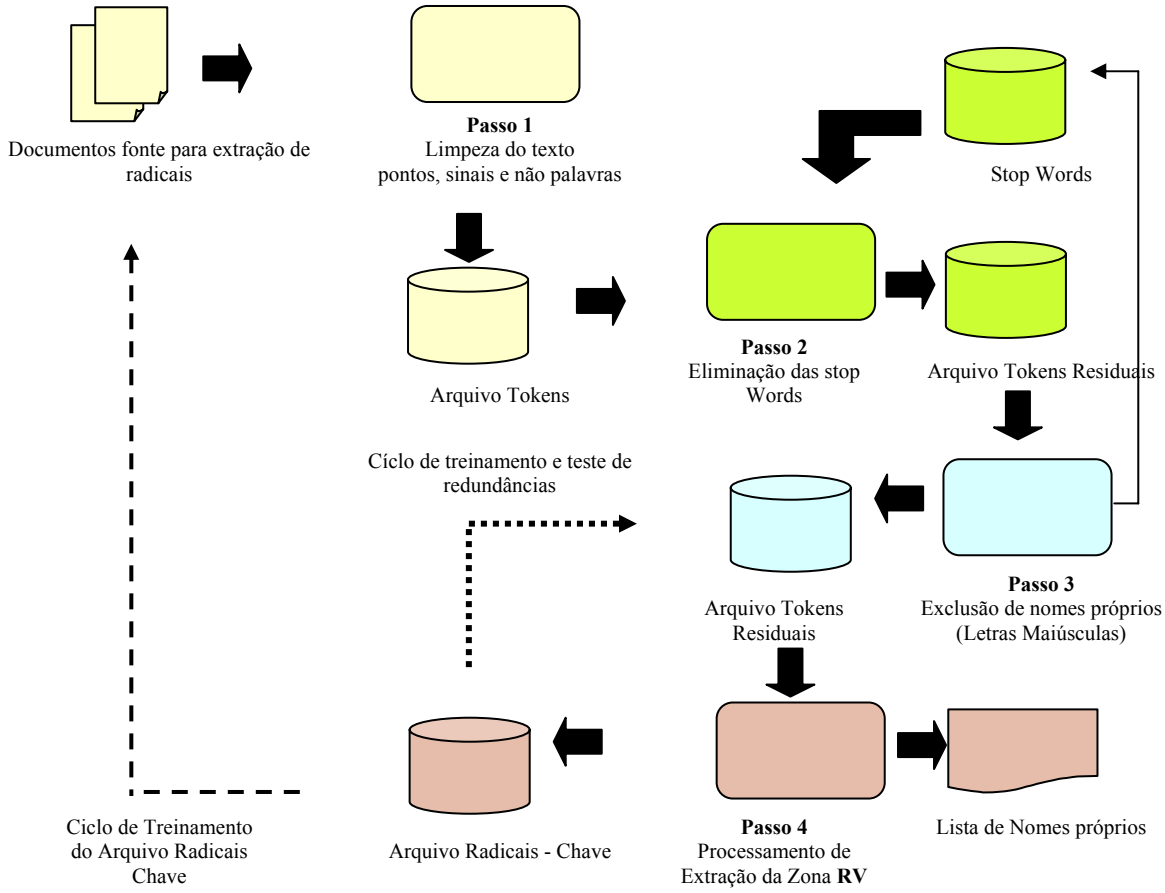


Figura 3.4 – Algoritmo Anaphora RV de extração de Radicais

3.4.2
Anaphora PCh para extração de palavras-chave

A extração de palavras-chave está baseada nos conteúdos da estrutura de radicais extraídos pelo algoritmo Anaphora RV. Nesta fase são usados textos policiais típicos, de onde são pesquisadas palavras-chave derivadas dos radicais.

O dicionário de radicais é uma estrutura temporária, criada para desenvolvimento do dicionário especialista.

Diferentemente do tratamento executado pelo algoritmo **Anaphora RV,** onde palavras como “crime”, “criminoso” ou “criminal” são reduzidos ao seu formato simplificado “crim”, na construção do dicionário especialista a flexão de gênero (masculino e feminino) e número (singular e plural) é significativa para modelagem da rede semântica. Palavras como “criminoso” e “criminosa” representam diferenças fundamentais na identificação de atores envolvidos, com

implicações nos relacionamentos derivados dos conceitos extraídos para construção do dicionário especialista.

A extração das palavras-chave é executada pelo algoritmo **Anaphora PCh** baseado no algoritimo **KEA,** que extrai e treina palavras-chave (Witten et al.,1999 - (b)).

O algoritmo está baseado no tratamento de palavras-chave existentes e documentos fonte típicos para treinamento e especialização. O principal objetivo do algoritmo é construir uma estrutura que forneça apoio à modelagem de redes semânticas, onde os conceitos extraídos representam os nós da rede e os relacionamentos entre os conceitos representam os seus arcos de conexão ou vértices. O algoritmo é executado em quatro passos.

- **Passo-1 -** Limpeza do Texto e Geração de Tokens

  O primeiro passo seleciona Tokens do texto de treinamento, extraindo todos sinais, pontuação e não-palavras.
- **PCh Passo 2** - Eliminação dos Tokens Inúteis (Stop Words)

  Os Tokens residuais são confrontados contra uma lista de Stop-words, eliminando-se aqueles, porventura, presentes nesta lista.
- **PCh Passo 3 -** Exclusão das Maiúsculas

  Uma análise sintática é executada para identificação de palavras Maiúsculas simples e maiúsculas combinadas (conectadas pelas preposições *da, de, do*). Estas palavras são eliminadas da lista de Tokens selecionados para o dicionário especialista, e formam uma lista de candidatos potenciais para o dicionário de nomes próprios ou de Stop-words
- **PCh Passo 4 -** Atualização do Dicionário Especialista

  Atualiza o dicionário especialista eliminando as redundâncias. Uma lista adicional de nomes próprios e Stop-words não contidas nos dicionários de apoio é gerada para atualização,
- **PCh Fim -** Reciclagem

  Os arquivos obtidos como produto do algoritmo **Anaphora PCh** retornam para novas inclusões de palavras-chave, como parte do treinamento e especialização de conteúdos. Este ciclo se encerra quando os conteúdos dos arquivos de apoio são considerados eficientes para extrações de

conceitos e modelagem de redes de relacionamento com mínimas perdas ou erros de representação.

O fluxo apresentado na Figura 3.5 apresenta os passos para desenvolvimento do algoritmo **Anaphora PCh**.

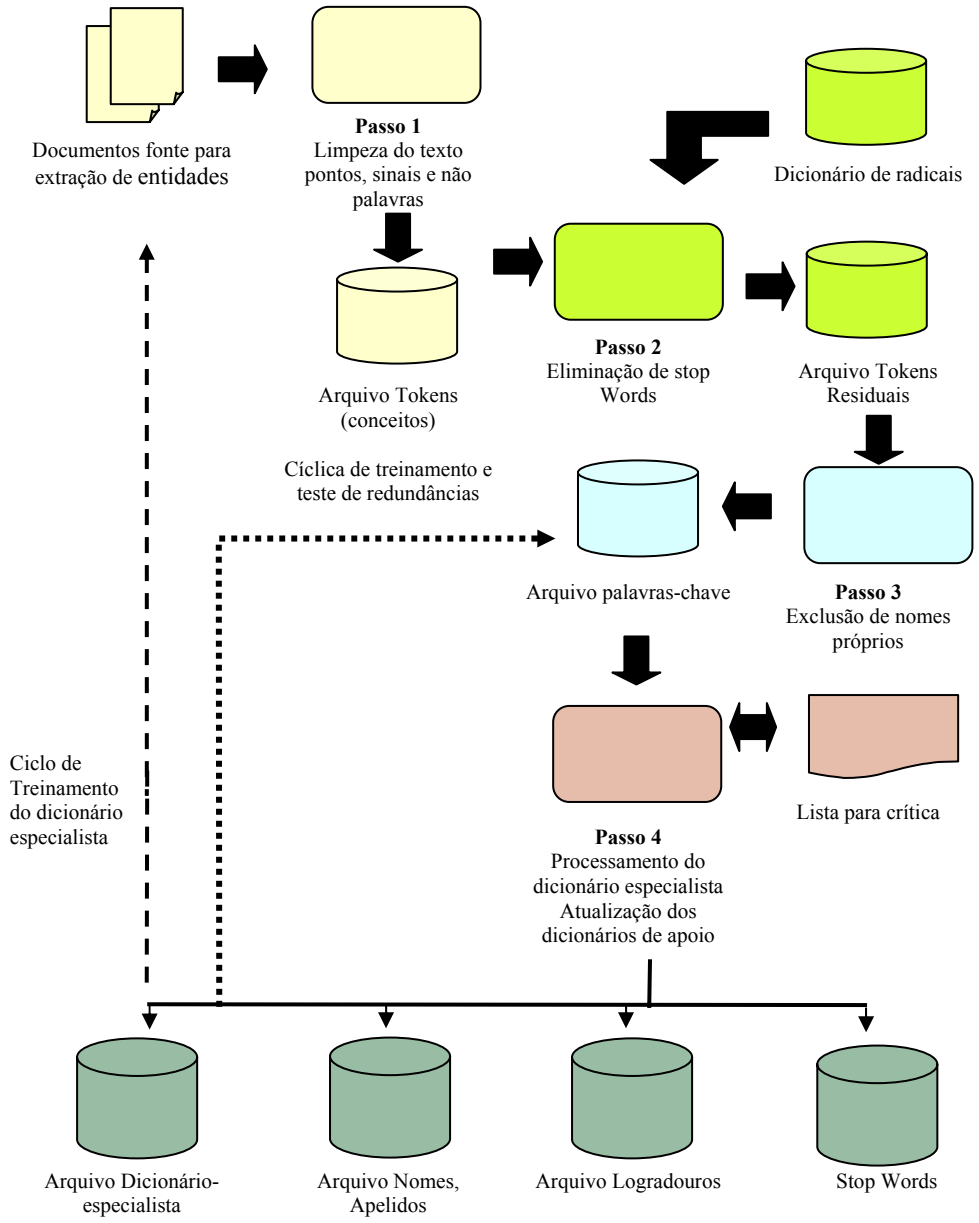


Figura 3.5 – Algoritmo Anaphora PCh para construção do Dicionário Especialista

Cada ciclo de processamento de palavras-chave trata a extração de forma independente, identificando redundâncias e stop-words, através de contínuas interações com os documentos de treinamento.

As estruturas são submetidas a uma crítica manual (após o encerramento da fase de treinamento). A especialização contínua dos dicionários minimiza gradualmente a fase de inspeção manual dos novos dados extraídos.

Esta crítica tem uma forte influência sobre a lista de palavras negativas pré-selecionadas. Devido aos erros comuns cometidos na transcrição das ocorrências policiais, diversos ruídos podem ser introduzidos na geração dos *tokens*. A crítica dos dados extraídos auxilia a depuração e conferência dos radicais selecionados de forma automática pelo algoritmo.

A especialização do dicionário é alcançado após diversas interações, executadas no conjunto selecionado para o treinamento da estrutura de palavras-chave.

O ciclo **Texto → Radicais → Dicionário_Radicais → Texto → Radicais → Dicionário_Radicais** irá especializando o dicionário para identificação genérica de entidades úteis em documentos policiais.

Após o encerramento da fase de treinamento, as palavras-chave selecionadas para o dicionário especialista serão manualmente classificadas em três classes de entidades. Estas classes relacionam-se com o nível de detalhes que serão extraídos dos documentos sob forma de entidades úteis. As classes correspondem aos tipos de entidades que serão usadas no apoio à extração, desde as mais específicas como nome de facções até as mais genéricas, como viaturas, armas e drogas. As entidades mais específicas produzem um menor volume de entidades, enquanto as mais genéricas podem produzir grandes volumes de entidades, devendo ser empregadas no tratamento de pequenos volumes de documentos.

As três classes de entidades do dicionário Especialista foram assim denominadas:

- Tipo 1 - palavras-chave para um tratamento com baixo nível de detalhes extraídos.
- Tipo 2 - palavras-chave para um tratamento com médio nível de detalhes extraídos.
- Tipo 3 - palavras-chave para um tratamento com grande nível de detalhes extraídos.

3.5
Sistemas para apoio à extração de dicionários

Este tópico apresenta os sistemas desenvolvidos para construção dos dicionários especialistas e descreve os testes executados com os sistemas empregados para construção dos dicionários descritos no método proposto nesta Tese.

Foram desenvolvidos dois aplicativos usados nas fases de construção do Dicionário Especialista: o Sistema **Stemmer Anaphora** e o sistema **DicTools**. Estes aplicativos são responsáveis pela execução do algoritmo Anaphora RV, utilizado para extração de radicais e do algoritmo Anaphora PCh, utilizado para extração de palavras-Chave e construção do Dicionário Especialista.

3.5.1
Descrição do Sistema Stemmer Anaphora

O Sistema Stemmer Anaphora é um aplicativo para suporte às funções de extração de radicais (stemmeer) de documento textuais. A principal inovação oferecida pelo aplicativo é a sua capacidade de suportar múltiplos algoritmos e permitir alterações do algoritmo em tempo de operação do sistema, introduzindo parâmetros que atualizam regras e exceções incorporadas ao algoritmo ativado.

As regras e exceções são introduzidas ou alteradas através do editor Bloco de Notas do Windows© ou qualquer outro aplicativo editor de textos disponível.

Um arquivo de extensão **AGO**, tipo texto, contendo as referências do algoritmo é lido na fase de inicialização. Este arquivo orienta o algoritmo ativado para a seqüência do tratamento de radicais a ser seguido. Cada passo do algoritmo está associada a uma linha do arquivo texto AGO. Para cada linha com a especificação inicial "**Regra**" (passo do algoritmo) está associado um arquivo contendo regras específicas que comandarão o tratamento do radical. Por exemplo, na regra para o feminino, a terminação **ona** deverá ser convertida para a terminação **ão**.

A Tabela 3.6 apresenta as declarações dos passos associados ao algoritmo Orengo e Huyck registrados em um arquivo **AGO**.

O caracter **@** atua como delimitador dos parâmetros lidos.

Cada passo está representado por uma linha de parâmetros no arquivo AGO, indexando regras e exceções armazenadas em arquivos associados. Estas indexações servem como orientação do aplicativo para carga das regras do algoritmo no momento de sua inicialização.

Cada passo, indicado pela palavra "Regra" no início da linha, contém o formato seguinte:

Regra@02@Feminino@exceçõesfeminino.txt@regrasfeminino.txt

- declaração que a linha é um passo do algoritmo: **Regra**
- radical mínimo após a extração: **02**
- nome da regra: **Feminino**
- nome do arquivo associado às exceções: **exceçõesfeminino.txt**
- nome do arquivo associado às regras: **regrasfeminino.txt**

Tabela 3.6 – Regras associadas ao algoritmo Orengo e Huyck .

Parâmetros	Conteúdo do arquivo
Título do algoritmo	Titulo@Orengo & Huyck
Numero de Passos	Passos@8
Regra	Regra@02@Plural@exceçõesplural.txt@regrasplural.txt
Regra	Regra@02@Feminino@exceçõesfeminino.txt, regrasfeminino.txt
Regra	Regra@03@Advérbio@exceçõesadverbio.txt@regrasadverbio.txt
Regra	Regra@02@Aum/Dimin@exceçõesAumentativoDiminutivo.txt,Regras AumentativoDiminutivo.txt
Regra	Regra@02@Substantivo@exceçõessubstantivo.txt@regrassubstantivo.txt
Regra	Regra@02@Verbo@exceçõesverbos.txt@regrasverbos.txt
Regra	Regra@02@Vogal@exceçõesvogal.txt@regrasvogal.txt
Regra	Regra@02@Acento@exceçõesacentos.txt@regrasacentos.txt

Fonte: o Autor

Os exemplos apresentados nas Figuras 3.6 e 3.7 demonstram a edição de regras e exceções aplicáveis ao Algoritmo Orengo e Huyck para o passo de redução do feminino, opções do algoritmo para redução do plural, feminino, advérbio, aumentativo / diminutivo, substantivo, verbo, vogal e acento.

Através de um editor de texto é possível introduzir uma nova regra ou uma nova exceção, substituir uma regra ou exceção existentes ou simplesmente observar regras e exceções armazenadas, sem alterar o conteúdo original dos

arquivos. Desta forma, podemos modelar um novo algoritmo com simplicidade, dispensando conhecimentos prévios de programação e dispensando re-compilação do código original do programa após alterações na estrutura do algoritmo.

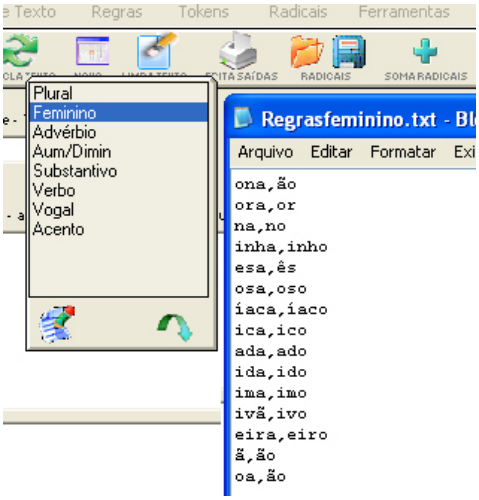


Figura 3.6 – Exemplo de regras, aplicadas ao algoritmo Orengo e Huyck, aplicativo Anaphora

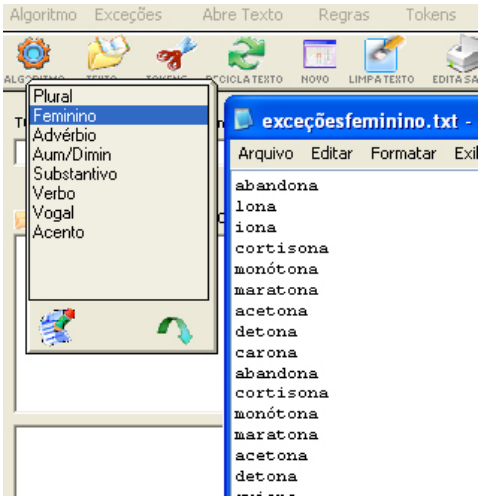


Figura 3.7 – Exemplo de exceções aplicadas ao algoritmo Orengo e Huyck, aplicativo Anaphora

Dois algoritmos foram incorporados ao Sistema Stemmer Anaphora: o algoritmo Anaphora RV, que trabalha com a região RV das palavras processadas e algoritmo Orengo e Huyck, desenvolvido para extração de radicais de palavras em português.

O algoritmo Orengo e Huyck foi traduzido para a linguagem de parâmetros suportada pelo sistema Stemmer Anaphora, recebendo contribuições não previstas no algoritmo original, como exceções de terminais aumentativos (sufixos em **ão** ou **ona**, conforme encontrados em vulcão e mamona).

Um problema de tempo para execução do aplicativo foi detectado, particularmente quando o sistema defrontava-se com uma grande relação entre o volume de radicais tratados e o tamanho do texto processado. O problema tinha o seu foco na pesquisa de palavras descartáveis, cujo acesso do algoritmo extrator de radicais exigia uma alta freqüência de acesso à lista de stop-words.

O conceito de grandes volumes aqui considera quantidades superiores a 1000 radicais e/ou textos superiores a 250.000 palavras.

Para permitir um melhor desempenho na pesquisa de stop-words, foi desenvolvido um algoritmo de acesso através de uma árvore *hashcode*[4], com a pesquisa primária baseada na letra inicial da stop_word.

A seguinte heurística foi implementada visando reduzir o tempo de pesquisa as stop-words:

- Segmentar as stop-words em listas parciais, organizadas pela letra inicial das palavras que passa a funcionar como chave de acesso às listas.
- Reduzir o tempo de acesso da pesquisa da palavra-chave às listas indexadas em ordem alfabética.

Esta heurística permitiu um melhor desenvolvimento na extração das raízes pesquisadas, cujo processamento para cada stop-words encontra-se na razão de 1:n consultas, onde **n** corresponde à uma sub-lista de stop-words iniciando com a letra da palavra pesquisada.

Uma vez que a árvore aponta diretamente para sub-listas organizadas pelas letras do alfabeto, a pesquisa funciona de forma seletiva, mais eficiente e rápida que uma pesquisa realizada usando um método de varredura seqüencial.

A Figura 3.8 apresenta o painel principal de funções do Sistema Stemmer Anaphora.

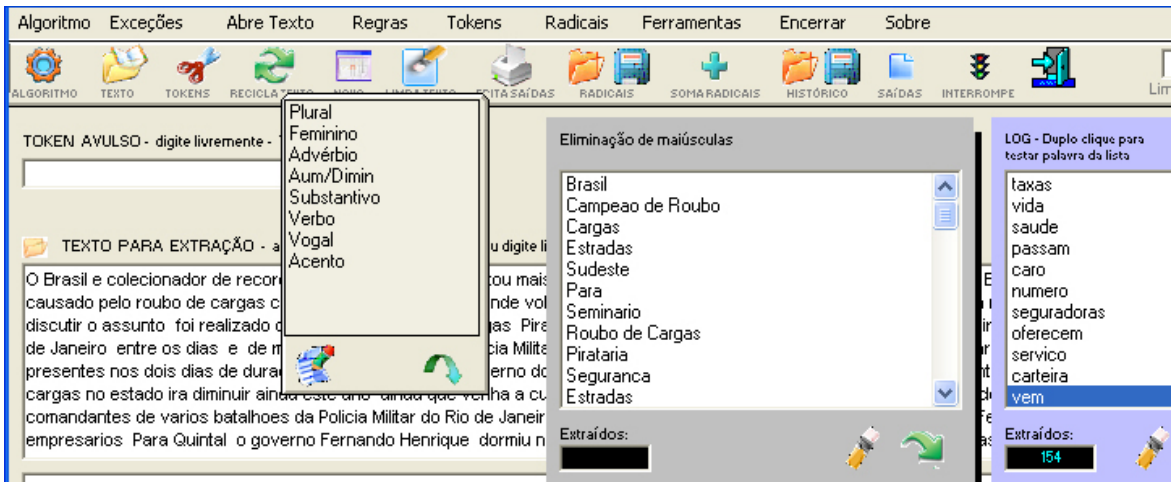


Figura 3.8 – Painel do Sistema Stemmer Anaphora

O Sistema Stemmer Anaphora oferece uma janela contendo os passos de evolução do algoritmo para cada palavra tratada e um relatório impresso opcional

[4] Hash code - um código único para referência de itens contidos em uma estrutura

que fornece um histórico com o registro da trilha de redução dos radicais tratados, permitindo o acompanhamento dos passos executados pelo algoritmo ativado.

A Figura 3.9 apresenta a trilha de redução do algoritmo Orengo e Huyck e a Figura 3.10 apresenta a trilha de redução do algoritmo Anaphora RV.

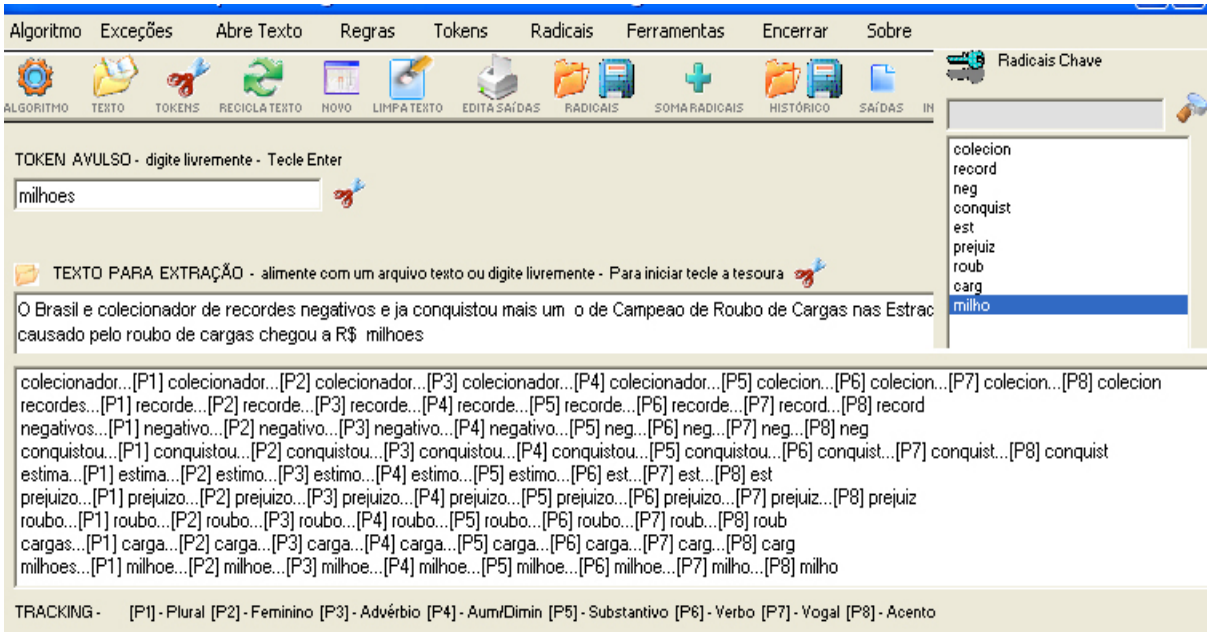


Figura 3.9 – Seqüência de redução de palavras- algoritmo Orengo e Huyck

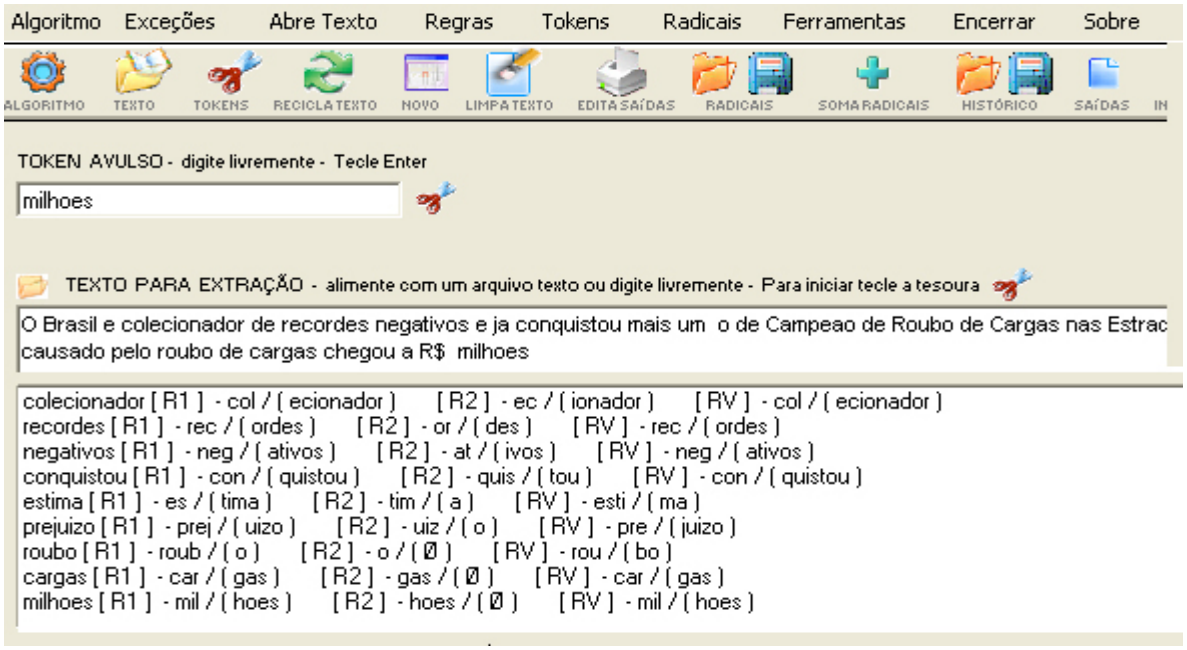


Figura 3.10 – Seqüência de redução de palavras- algoritmo Anaphora RV

A função [**soma radicais**] permite aumentar uma estrutura existente de radicais através de sucessivas interações do algoritmo, acrescentando-se à

estrutura anterior o resultado obtido com uma nova interação. Através deste recurso é possível treinar uma lista temática de radicais de forma recursiva. A depuração de duplicidades é executada em tempo de processamento, obtendo-se gradualmente uma lista especializada no tema desejado.

A Figura 3.11 apresenta a seqüência de treinamento da estrutura de radicais no Sistema Anaphora.

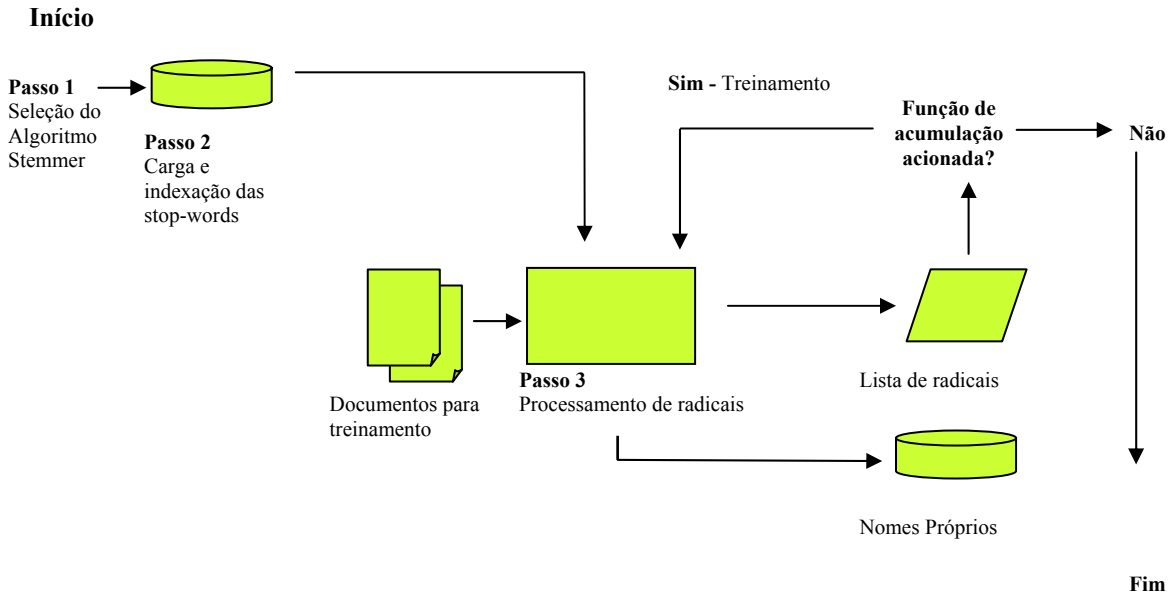


Figura 3.11 – Seqüência de treinamento de radicais no Sistema **Anaphora**

A programação do **Sistema Stemmer Anaphora**, destinado ao processamento de algoritmos **extratores de radicais** é inédita, tendo sido desenvolvido pelo autor em 2007, para os laboratórios requeridos pela Tese, tendo sido utilizada linguagem de programação Microsoft Visual Basic 6 e suportando alimentação de arquivos em formato texto (.txt), arquivos em formato texto-rico (.rtf) e arquivos preparados em Microsoft Word (.doc).

3.5.2
Sistema DicTools

O segundo aplicativo usado na seqüência de construção do dicionário especialista é o sistema proprietário DicTools que opera o algoritmo Anaphora PCh, usado para extração de palavras-chave de documentos textuais em

português. O sistema DicTools alimenta-se de listas de radicais treinados para extração de palavras-chave e utiliza o suporte de stop-words para eliminação de conceitos indesejáveis.

O Sistema Dictools também possui funções para extração de entidades úteis de documentos textuais, destinadas à modelagem de redes semânticas, orientado por dicionários eletrônicos de apoio, como dicionário especialista, dicionário de logradouros e dicionário de nomes próprios.

O sistema Dictools extrai *Tokens* de um texto digital, realizando uma crítica prévia da entrada, que compreende a análise léxica da palavra, teste de existência no dicionário especialista (duplicidade) e teste de exclusão no arquivo de stop-words. As palavras selecionadas são exibidas em uma lista de resultados, no final da interação. Esta lista, classificada na ordem inversa da freqüência de ocorrências processadas, apresenta a participação de cada palavra-chave selecionada no texto candidata à inclusão no arquivo dicionário especialista. O comando de inclusão do novo grupo de palavras-chave processadas para o dicionário especialista pode realizar-se de forma automática pelo sistema ou de forma manual, mediante confirmação do operador. Este ciclo de inclusões irá produzir a especialidade do dicionário, até que novas inclusões promovidas por treinamento revelem nenhum ou baixo nível de inclusões no dicionário especializado.

A consistência do dicionário especialista é mensurada através da metodologia seguinte:

- Uma coleção de documentos típicos é selecionada para aferição do modelo;
- É processada uma extração de entidades úteis, com base na estrutura de palavras chave do dicionário especialista;
- O indice **Revocação** é computado pela relação entre o volume de extrações de palavras-chave realizadas pelo sistema e o volume total de conceitos úteis existentes na coleção selecionada. Este indicador ira revelar a eficiência do treinamento, cujo valor desejavelmente deverá situar-se próximo a 1 (100% de eficiência).

A Figura 3.12 apresenta o painel principal de funções do Sistema DicTools.

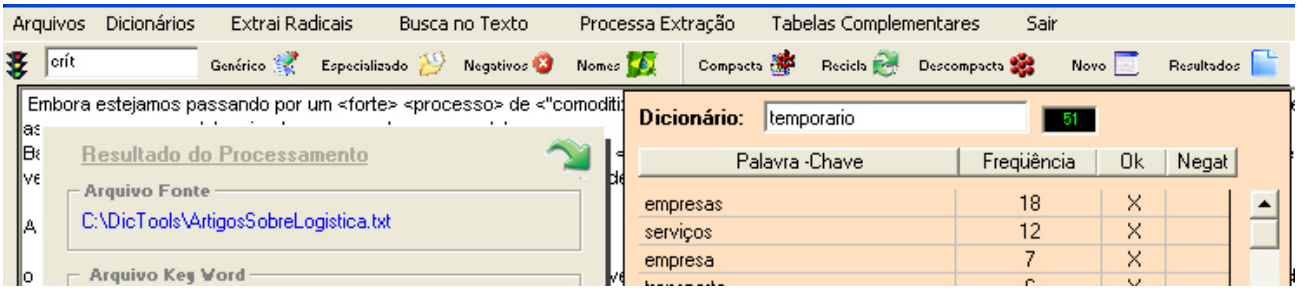


Figura 3.12 – Painel do Sistema DicTools

3.6
Avaliação do método para extração de palavras chave

Neste tópico são apresentados testes desenvolvidos para avaliação do método proposto para extração de palavras chave para construção de um dicionário especialista.

3.6.1
Comparação entre o método manual e automático

Um experimento foi desenvolvido para avaliação de resultados obtidos com a extração automatizada de palavras-chave. Foram selecionados boletins policiais ricos em informações criminais, totalizando dois anos de investigações sobre grupo de traficantes de drogas atuantes em países da América do Sul.

O experimento foi desenvolvido comparando resultados obtidos pelo processo automatizado com resultados obtidos por processo manual.

Para o processamento automatizado foi utilizado os algoritmos descritos no método proposto nesta Tese operados pelos aplicativos Anaphora e Dictools.

Três policiais experientes foram convidados para extrair manualmente entidades, usando a mesma coleção de históricos criminais processados pelo sistema automatizado.

A extração manual deveria seguir as mesmas regras da extração quanto à utilidade das entidades extraídas e aplicando as mesmas regras de rejeição das entidades não relevantes (stop-words).

Cada entidade deveria ser processada qualitativa e quantitativamente.

Entidades formadas por palavras compostas como *"chefe da quadrilha"* ou *"Venda de Drogas"* deveriam constar como apenas como uma única entidade.

Preposições como "*para*" ou "*desde*", palavras como "*enquanto*" ou "*principalmente*" não deveriam ser consideradas como entidades relevantes. Deveria haver concordância entre os três participantes quanto à inclusão ou rejeição das entidades tratadas no experimento.

Os resultados gerais deveriam, no final da extração manual, serem consolidados para fins de comparação com as entidades extraídas pelo sistema automatizado.

Os resultados apurados no experimento estão apresentados na Tabela 3.7 abaixo.

Tabela 3.7 – Resultados apurados no experimento da comparação entre Algoritmo automatizado e extração manual

Processo	**Entidades relevantes**			**Entidades não relevantes (Stop-words)**			**Total**
	Corretas	**Erradas (M-A)**	**Perdas (E) / (M)**	**Corretas**	**Erradas (N-B)**	**Perdas (C) / (N)**	
Automatizado	1884 (**A**)	177 **(E)**	8,5%	1203 (**B**)	76 **(C)**	5,9%	3340
Manual	2061 (**M**)	-		1279 (**N**)	-		3300
Relação (automatizado / manual)	0,91	-		0,94	-		

Fonte: o autor

(**A**) Entidades relevantes recuperadas pelo sistema

(**M**) Entidades relevantes recuperadas manualmente

(**E**) Entidades relevantes não recuperadas (**M - A**)

(**B**) Entidades não relevantes recuperadas

(**N**) Entidades não relevantes recuperadas manualmente

(**C**) Entidades não relevantes não recuperadas - (**N - B**)

Total de itens recuperados = (A+B)

Total de itens relevantes = (A+E)

Para avaliação dos resultados foram usados indicadores de Revocação e Precisão (Paralic & Smatana, 2005).

- Revocação mede a proporção de entidades relevantes recuperadas
- Precisão mede quantas entidades relevantes foram recuperadas

A **Revocação** refere-se à relação entre o número de entidades relevantes recuperadas pelo sistema e o número total de entidades existentes.

Revocação = Relevantes Recuperadas / Total de Itens Relevantes

A/(A+E) = 91,2

A **Precisão** ou relevância refere-se à qualidade da recuperação, ou seja, a relação entre o número de entidades relevantes recuperadas e o total de entidades relevantes e não relevantes recuperadas.

Precisão = Relevantes Recuperadas / Total Recuperadas

A/(A+B) = 60,3

No processamento automatizado foram extraídos 1884 entidades relevantes, comparando com 2021 entidades obtidas pelo processo manual, revelando uma relação de 91,2% entre o sistema automatizado e o sistema manual.

Para a pesquisa de entidades não relevantes, o processamento automatizado extraiu 1203 entidades não relevantes contra 1279 obtidas no processamento manual, obtendo uma relação de 94,2% entre o sistema automatizado e o sistema manual.

O índice geral para reconhecimento de entidades apresentou uma eficiência de 92,2% para identificação geral de entidades em textos livres.

3.6.2
Testes Comparativos entre algoritmos extratores de Radicais

Para fundamentar o processo de seleção do algoritmo Stemmer foi desenvolvida uma experimentação para avaliar dois tipos de algoritmos extratores de radicais, Foram realizados testes comparativos entre os algoritmos Anaphora RV (extrator de zonas) e algoritmo Orengo e Huyck modificado (extrator de sufixos).

Os resultados dos testes entre os algoritmos foram desenvolvidos observando a eficiência no tocante à capacidade de reconhecimento e extração de palavras derivadas e válidas.

A experimentação foi executada utilizando um conjunto de documentos de ocorrências policiais registrados no setor de inteligência da Polícia Civil do Estado de Rio de Janeiro (Fotocrim-Sinpol, 2003).

O conjunto de documentos selecionado para a experimentação apresentou um volume de 6825 palavras, dentre substantivos, nomes próprios, artigos etc.

O objetivo do experimento seria verificar a eficiência de extração de palavras-chave submetidas a ambos os conjuntos de radicais produzidos pelos algoritmos.

Os documentos foram submetidos aos algoritmo selecionados para o teste comparativo Anaphora RV e Orengo e Huyck, extratores de Radicais, usando o aplicativo Anaphora.

As estruturas resultantes das extrações de radicais foram então aplicadas na extração de palavras-chave usando um novo conjunto de documentos para aferição de resultados. Para tratamento das palavras chave foram utilizados o aplicativo DocTools e o algoritmo Anaphora PCh.

A Tabela 3.8 apresenta uma comparação de resultados para extração de radicais, aplicando-se os algoritmos Orengo e Huyck (2001) e Anaphora RV, observando-se a eficiência dos algoritmos selecionados.

Tabela 3.8 – Resultados do teste comparativo entre algoritmos Orengo e Huyck e Anaphora RV

	Algoritmo RV	**Algoritmo Orengo e Huyck**
Radicais Extraídos (a)	529	799
Conceitos totais extraídos, aplicando-se os radicais extraídos (b)	859	656
Índice de geração de palavras derivadas, por radical (a/b)	1,6 : 1	1,2 : 1
Eliminação de palavras inúteis através de Stop Word	60	51
Aproveitados (c)	799	605
Índice acertos (c/b)	93%	92%

Fonte: o autor

Algoritmo RV gerou 30% adicionais de palavras derivadas sobre a produção de palavras geradas pelo algoritmo **Orengo e Huyck,** utilizando um volume 34% inferior de radicais.

Algoritmo RV apresentou uma relação média de 1 radical para 1**,**6 palavras derivadas, contra uma relação média de 1 radical para 1**,**2 palavras derivadas pelo Algoritmo **Orengo e Huyck** (relação 33% superior).

Algoritmo RV eliminou 60 palavras inúteis contra 51 palavras inúteis eliminados pelo algoritmo **Orengo e Huyck**.

O volume maior de eliminações de stop-words processado pelo **Algoritmo RV** deve-se à tendência do algoritmo RV em produzir um maior volume de cognatos derivados, característico dos radicais reduzidos. Este processo é denominado de "falsos amigos" (Krovetz, 1997), quando cognatos de diferentes famílias são extraídos tendo como origem radicais homônimos.

Listas amplas de stop words devem estar incorporadas ao processamento do algoritmo **RV** que utiliza radicais reduzidos, com maior grau de derivação. Quanto maior for o volume de cognatos, menos geral será o significado da palavra (Monteiro, 1986).

Por exemplo, **general** (patente) e **generoso** convergem para o mesmo radical RV, [*gen*]. Seria necessário considerar "generoso" como stop-word, que seria usado para eliminação de palavras inválidas no passo de verificação das Stop-words.

O índice percentual de aproveitamento referente a acertos de **RV** foi de 93%, contra 92% de **OH**, denotando uma igual capacidade referencial de acertos entre os algoritmos.

O resultado obtido no estudo comparativo entre os dois métodos ponderou favoravelmente na escolha do algoritmo Anaphora RV como mais recomendável para desenvolvimento do dicionário especialista. Consideramos os princípios seguintes para escolha do algoritmo Anaphora RV:

a) O Algoritmo para Remoção de sufixos de Orengo de Huyck apresenta tendência *Understemming (*sufixo não é removido completamente) reduzindo o grau do semantema extraído, podendo perder derivados (Chaves, 2003).
b) A tendência destrutiva, observada em ocorrências de *Overstemming* (parte do stem[5] é suprimido) pode causar perda eventual de radicais e, conseqüentemente, de todos os cognatos derivados (Chaves, 2003).
c) O algoritmo Anaphora RV não produz perdas causadas pela destruição de parte do stem, conservando o radical reduzido em uma dimensão mínima, que pode variar de dois a quatro caracteres.
d) O algoritmo de tratamento da região RV produz radicais curtos e de maior grau morfológico, mantendo as propriedades derivacionais não destrutivas.
e) O algoritmo de tratamento da região RV apresenta capacidade maior de geração de derivados, inerente de radicais com maior grau morfológico.

[5] Stem = raiz da palavra

3.6.3
Extração e treinamento de entidades para construção de um dicionário especialista temático

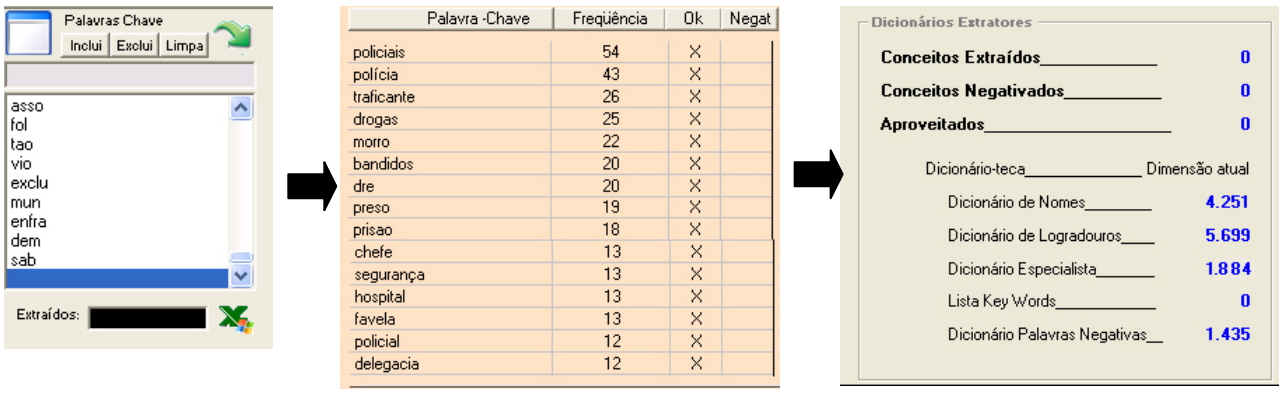
O índice de revocação obtido na avaliação do método para extração de palavras chave descrito no tópico 3.6.1 (91,2%) demonstrou a capacitação e eficiência do método proposto nesta Tese para desenvolvimento de um dicionário especializado, orientado para o domínio textual compreendido pela coleção de históricos policiais de referência.

Para construção de um dicionário especialista de apoio ao método proposto nesta Tese foram selecionados dois conjuntos de documentos ao acaso, extraídos da coleção de documentos da Base Criminal Fotocrim-Sinpol (2003). O primeiro dos conjuntos selecionados contendo 60 documentos foi usado para extração e treinamento de uma estrutura contendo palavras-chave e o segundo conjunto contendo 60 documentos foi usado para aferição da estrutura treinada.

Os resultados do treinamento para extração de palavras-chave foram os seguintes:

- Palavras-chave extraídas: 1884 (Perdas - 5,5%, Erros - 2,4%)
- Nomes próprios selecionados: 377
- Tokens selecionados: 10.009

A Figura 3.13 apresenta um ciclo preliminar da extração de palavras-chave, ilustrando os resultados parciais de um ciclo da extração.



Radicais Palavras-chave Resumo

Figura 3.13 – Resumo de um ciclo de processamento para extração de palavras-chave

O treinamento foi executado em 20 ciclos, até que nenhum token residual foi gerado no último ciclo de treinamento. A estrutura de palavras-chave foi a mesma utilizada na fase 1, durante o processamento de extração de radicais. Após a aplicação da coleção de documentos reservados para aferição de palavras-chave extraídas a estrutura foi considerada suficientemente treinada para aplicação na modelagem de Mapas de Inteligência.

3.6.4
Dicionários Complementares de Apoio à Extração

Três dicionários obtidos através de método manual e semi-automático, complementam a estrutura de dicionários do sistema de dicionários de apoio à extração (Lee,1998 ; Witten et al.,1999).

- **Dicionário de Logradouros**

  Contém 5699 referências geográficas comuns utilizadas em boletins de ocorrências policiais do Estado do Rio de Janeiro, como Siglas, Municípios, Estados. O **Dicionário de Logradouros** importou dados do sistema de informações do IBGE (1997).

- **Dicionário de Nomes e Apelidos**

  Contém 4251 referências nominais e apelidos utilizadas em boletins de ocorrências policiais do Estado do Rio de Janeiro. O dicionário **Nomes** e **Apelidos** importou dados do sistema de inteligência Fotocrim- Sinpol, Polícia Civil, Rio de Janeiro (2003).

- **Dicionário de Palavras Negativas (Stop-words)**

  Contém 1435 palavras, obtidas através de regras, operação manual e contínuo treinamento. Destina-se à remoção de termos freqüentes, considerados inúteis, através da comparação do conceito extraído contra uma lista de "palavras-negativas" que devem ser removidas (Blum & Langley, 1997; Sanderson, 2003; Rijsbergen, 2007).

3.7
Conclusões deste Capítulo

Este capítulo demonstrou um método para construção de um dicionário temático especializado em linguagem característica de históricos policiais.

Inicialmente radicais foram extraídos de textos livres e treinados com auxílio de algoritmos e ferramentas sistêmicas. As estruturas treinadas foram aplicadas como apoio à extração de palavras-chave em coleções de históricos policiais e utilizadas para construção de um dicionário especializado temático.

O dicionário tornou-se especializado através de treinamento e reciclagem de seu conteúdo até um estágio considerado como eficiente para apoio à extração de entidades úteis em históricos policiais. As entidades extraídas serão posteriormente utilizadas para modelagem de Mapas de Inteligência, onde serão aplicadas ferramentas de análise para extração de conhecimento em investigações policiais.