EXTRAÇÃO DE RELAÇÕES SEMÂNTICAS EM REIVINDICAÇÕES DE PATENTES

Danilo Silva de Carvalho

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Rio de Janeiro
Novembro de 2014

EXTRAÇÃO DE RELAÇÕES SEMÂNTICAS EM REIVINDICAÇÕES DE PATENTES

Danilo Silva de Carvalho

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____
Prof. Felipe Maia Galvão França, Ph.D.

_____
Prof. Priscila Machado Vieira Lima, Ph.D.

_____
Prof. Geraldo Bonorino Xexéo, D.Sc.

_____
Prof. João Carlos Pereira da Silva, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
NOVEMBRO DE 2014

Carvalho, Danilo Silva de

Extração de Relações Semânticas em Reivindicações de Patentes/Danilo Silva de Carvalho. – Rio de Janeiro: UFRJ/COPPE, 2014.

XVIII, 141 p.: il.; 29, 7cm.

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2014.

Referências Bibliográficas: p. 77 – 85.

1. Extração de Relações Semânticas. 2. Processamento de Linguagem Natural. 3. Aprendizado de Máquina. I. França, Felipe Maia Galvão *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A todos os professores, que com sua dedicação ajudam a mover e transformar aquilo que a humanidade mais precisa: a sabedoria. Com ela, superamos as dificuldades passadas, presentes e futuras.

Agradecimentos

Gostaria de agradecer em primeiro lugar aos meus pais, que me ofereceram bons exemplos e educação, sem os quais eu não teria chegado a este ponto. Embora hoje distantes de minha vida acadêmica, devo a eles todas as principais correções de rumo que precisei até me tornar um adulto.

Aos meus familiares, em especial às minha avós que infelizmente não estão mais entre nós, uma por seu grande apoio moral e muitas vezes financeiro desde o meu nascimento, e a outra por sempre me lembrar de como enfrentar dificuldades com bom humor. Estou certo de que estariam muito satisfeitas com minhas conquistas.

Aos meus professores do ensino fundamental, em especial à profª Margareth (1-2ª series), profª Lilian (4ª série), profs. Eldamir e Simone (5ª série), profs. Zaíra e Hélio (6ª série). Eles não só me deram os alicerces para a construção do conhecimento de que disponho hoje, mas me motivaram a examinar as diversas maneiras, ainda que conflitantes, com que este conhecimento pode ser compartilhado.

Aos meus professores do Ensino Médio e Técnico, em especial ao prof. Virgílio, por sua atenção e incentivo a exploração de técnicas e conteúdos muitas vezes fora do currículo escolar e por sua disposição em me mostrar seu trabalho e avanções acadêmicos. Também agradeço especialmente ao prof. José Marmute, por seus valiosos ensinamentos de programação, que definiram minha linha de atuação profissional nos anos que se seguiram, e pelo seu apoio na conquista do meu primeiro emprego, que tornou possível e proveitosa minha formação superior.

Ao meu médico, Dr. José Carlos Lino, sem o qual eu talvez não estivesse vivo para terminar este trabalho, e também por sua amizade e enriquecedoras conversas.

Aos meus professores do curso de Ciência da Computação da UFRJ, em especial aos profs. Milton Ramirez, Eber Schmitz, Gabriel Pereira, Mario Benevides, Geraldo Zimbrão, Geraldo Xexéo e João Carlos, por sua atenção especial e persistente e pelos conselhos valiosos, que ajudaram a solidificar meu interesse na carreira acadêmica. Novamente aos profs. Mario, Eber, João Carlos, e Xexéo por me incentivarem amplamente a cursar o mestrado.

Aos meus colegas de graduação e mestrado: Daniel Alves, Daniel Nunes, Diego Souza, Douglas Cardoso, Flávia Vieira, Hugo Carneiro, Israel Zinc, João Amarante, Kleber Aguiar, Leandro Marzulo, Leonardo, Marden, Paulo Brandt, Rafael Lima,

Roberta Lopes, Rodrigo Rodovalho, Saulo Oliveira, Taísa Martins, Vinicius Serva, e outros que minha memória deixou escapar. Todo o apoio que recebi deles e os bons e maus momentos que passamos jamais serão esquecidos.

Aos meus colegas de pesquisa, André Freitas, Bianca Pereira e Fabrício Firmino, pela grande amizade e apoio na aventura em que participamos na Irlanda. Nossas conversas e convivência me engrandeceram como pessoa e aprendiz de pesquisador.

Aos meus professores do mestrado Carlos Pedreira, Gerson Zaverucha, Geraldo Zimbrão e Sérgio Excel, pela motivação e confiança em minhas capacidades, e por me ajudarem a definir o rumo da minha pesquisa.

Aos meus orientadores: prof. Felipe França e profª Priscila Lima por me acolherem em seu núcleo de trabalho e principalmente pelo forte apoio acadêmico, moral e pela confiança depositada em mim em todas as nossas atividades. Sem tudo isso, este trabalho não seria possível.

Por fim, agradeço ao PESC, na figura de seus organizadores e colaboradores, por oferecer a oportunidade do excelente curso de mestrado que estou concluindo, e também à CAPES, CNPq e FAPERJ pelo apoio financeiro que mantém esta instituição funcionando.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

EXTRAÇÃO DE RELAÇÕES SEMÂNTICAS EM REIVINDICAÇÕES DE PATENTES

Danilo Silva de Carvalho

Novembro/2014

Orientadores: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Programa: Engenharia de Sistemas e Computação

Nos últimos anos, o foco econômico industrial em todo o mundo vem sendo desviado, passando da produção de ativos tangíveis para o conceito de Propriedade Intelectual, cuja proteção é regulamentada em muitos países pelo sistema de patentes. Com um crescente número de patentes concedidas, a gestão de informações relativas à inovação tornou-se uma tarefa árdua, levando ao desenvolvimento de diversas abordagens para sua automatização. Nestas abordagens predomina o uso de técnicas de Processamento de Linguagem Natural, mas as características deste tipo de documento criam grandes dificuldades para seu uso sem o auxílio de recursos externos, como ontologias de patentes, que limitam sua aplicação. Nesta dissertação é apresentado um método para extração de informações das reivindicações de patentes, através da identificação de unidades de significado relevantes aos documentos, na forma de fragmentos de texto chamados "segmentos semânticos". Este método utiliza apenas exemplos de reivindicações já segmentadas como ponto de partida para a extração, sendo portanto independente de outros recursos externos e aplicável a qualquer tipo de patente. A hipótese usada na condução do trabalho foi a de que há uma forte correlação entre a forma (sintaxe) e o significado em textos factuais, onde a ausência de ambiguidade é um requisito importante. Os experimentos conduzidos confirmaram tal hipótese, mostrando que é possível distinguir e relacionar uma parcela considerável das informações relevantes contidas nos documentos analisados. Os experimentos também mostraram que uma pequena quantidade de exemplos já é suficiente para a identificação das informações com maior regularidade na forma e que a abrangência das informações obtidas está positivamente relacionada à quantidade de exemplos apresentados.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

SEMANTIC RELATION EXTRACTION FROM PATENT CLAIMS

Danilo Silva de Carvalho

November/2014

Advisors: Felipe Maia Galvão França
Priscila Machado Vieira Lima

Department: Systems Engineering and Computer Science

In recent years, industrial economic focus has been changing all over the world, diverging from the production of tangible assets to the concept of Intellectual Property, for which protection is regulated by the patent system in many countries. With the increasing number of granted patents, the management of innovation related information has become a very difficult task, leading to the development of several approaches for its automation. In such approaches, the use of Natural Language Processing techniques is predominant, but characteristics of those documents impose considerable difficulties to the use of such techniques without the employment of external resources, such as patent ontologies, limiting their application. This dissertation presents a method for information extraction from patent claims, by the identification of relevant units of meaning for the documents, in the form of text fragments called "semantic segments". This method uses only examples of already segmented claims as the starting point for extraction, thus being independent from external resources and can be applied to any type of patent. The hypothesis adopted in the course of this work was that there is a strong correlation between the form (syntax) and the meaning on factual texts, where the absence of ambiguity is an important requirement. The experiments conducted confirmed such hypothesis, showing that it is possible to distinguish and relate a significant part of the relevant information in the analyzed documents. The experiments have also shown that a small number of examples is enough for identifying the information with the most regular forms, and that the recall of the information obtained is positively related to the number of examples presented.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ANN	*Artificial Neural Networks* - Redes Neurais Artificiais, p. 23
INPI	Instituto Nacional de Propriedade Industrial, p. 37
NER	*Named Entity Resolution/Recognition* - Resolução de Entidade Nomeadas, p. 2
OCR	*Optical Caracter Recognition* - Reconhecimento Óptico de Caracteres, p. 37
OWL	*Web Ontology Language* - Linguagem de Ontologias para a *Web*, p. 14
PCFG	*Probabilistic Context-Free Grammar* - Gramática Livre de Contexto Probabilística, p. 17
PDF	*Portable Document Format* - Formato Portátil de Documentos, p. 37
PLN	Processamento de Linguagem Natural, p. 2
POS	*Part-of-Speech* - Parte do Discurso: palavra de um texto, p. 16
RAM	*Ramdom Access Memory* - Memória de Acesso Aleatório, p. 2
RDF	*Resource Description Framework* - Arcabouço para Descrição de Recursos, p. 14
RIPPER	*Repeated Incremental Pruning to Produce Error Reduction* - Poda Incremental Repetitiva para Produzir Redução de Erro, p. 65
SVM	*Support Vector Machine* - Máquina de Vetor Suporte, p. 65
WANN	*Weightless Artificial Neural Networks* - Redes Neurais Artificiais Sem Peso, p. 25

WiSARD *Wilkie, Stonham & Aleksander's Recognition Device* - Dispositivo de Reconhecimento de Wilkie, Stonham & Aleksander, p. 25

Capítulo 1

Introdução

1.1 Motivação

Nos últimos anos, empresas e governos em todo o mundo têm participado de uma rápida transição de valores comerciais: de ativos tangíveis para o conceito de *Propriedade Intelectual*, com a regulamentação buscando seguir o ritmo de tal mudança. O desenvolvimento de esquemas e processos vem se tornando uma tarefa importante para os negócios e a academia. Contudo, o gerenciamento da informação relativa às inovações é uma tarefa árdua, que envolve a análise de uma vasta quantidade de documentos jurídicos e acadêmicos.

A *Recuperação de Informações em Patentes* é uma forma de facilitar tal tarefa, obtendo as partes mais relevantes dos documentos de patente, e.g., autor e assunto, e organizando-as em bases de conhecimento pesquisáveis para fácil acesso. Infelizmente, estes documentos são escritos predominantemente em linguagem natural, o que representa um grande desafio para a identificação correta das partes relevantes, especialmente termos novos ou inéditos. *Sistemas de Extração de Informação* oferecem uma solução para a representação de textos não estruturados como os em linguagem natural, analisando padrões terminológicos ou linguísticos. Isto é de grande relevância para os documentos de patente, para os quais a categorização e indexação estão em alta demanda, particularmente nas companhias de tecnologia, onde o gerenciamento de propriedade intelectual vem se tornando uma atividade crucial.

Algumas organizações têm tomado iniciativas no sentido de disponibilizar dados de patentes na internet, como a EPS [1] , epoline [2] (Europa) e Google's USPTO public downloads [3] (Estados Unidos), mas os formatos ainda não são consistentes e mui-

[1] European Publication Server (http://patentinfo.european-patent-office.org/off_pubs/pub_-serv/).

[2] http://www.epoline.org

[3] http://www.google.com/googlebooks/uspto.html

tos outros escritórios de patente ao redor do mundo não publicam seus documentos online ou o fazem em formatos não textuais, como é o caso do escritório de patentes brasileiro (INPI [4]), que publica formulários de papel "escaneados" em formato PDF. Portanto, um sistema robusto de extração de informações de patentes deve estar fundamentado em técnicas eficazes de *Processamento de Linguagem Natural* (PLN), para ser capaz de lidar com grande variedade de domínios de conhecimento e formatos. Desafios importantes em PLN para documentos de patente incluem: localização de termos, Reconhecimento e Resolução de Entidades Nomeadas (*Named Entity Recognition/Resolution* – NER, no inglês), análise sintática e extração de relações e funções semânticas.

Além da questão do acesso, os textos de patente possuem características únicas que dificultam o uso de técnicas tradicionais de PLN para extração de informações. Dentre estas, destacam-se a variedade de formatação dos documentos, i.e., como as seções são construídas e estão dispostas, e a predominância de sentenças longas e complexas, com múltiplas referências a outras partes do texto e a elementos não textuais como figuras.

Contudo, uma grande parcela das informações relevantes dos documentos de patente está concentrada na seção de reivindicações desses documentos. Elas contém a principal informação textual sobre o objeto de proteção legal, incluindo as referências para ilustrações e outros documentos. Além disso, são "bem comportadas" linguisticamente, em comparação com o resto do documento de patente, devido ao fato de serem o principal tópico de avaliação e litígio, e portanto são escritas de maneira a evitar ambiguidade ou inconsistência.

Neste trabalho é apresentado um método supervisionado para extração de informação semântica a partir de reivindicações de patente, usando estruturas sintáticas semanticamente anotadas, que são utilizadas para treinar um classificador neural baseado em RAM e obter conjuntos de informações frasais importantes, que são usados posteriormente para anotar reivindicações fora do conjunto de treinamento. A informação extraída tem a forma de sequências de palavras chamadas *segmentos semânticos*, organizadas em triplas *(sujeito, predicado, objeto)* para construção de grafos de relacionamento entre segmentos. O uso de reconhecimento de padrões visa contornar as dificuldades citadas, através da captura de informações implícitas nas estruturas sintáticas, evidenciadas pelas anotações semânticas. Desta forma, o método também dispensa o uso de ontologias de patentes, podendo o mapeamento ontológico ser feito após a extração dos segmentos.

[4]Instituto Nacional de Propriedade Industrial

1.2 Trabalhos relacionados

O trabalho de extração de informação em documentos de patente é caracterizado pela dificuldade em isolar o conhecimento de domínio do texto, dada a vasta cobertura de domínios apresentada pelas patentes. Portanto, este tipo de trabalho é frequentemente associado à pesquisa de ontologias superiores e de domínio, ambas na forma de extração de informação taxonômica e reconhecimento de entidades, e na forma de construção de ontologias a partir da extração de termos e relações. A Análise Semântica é geralmente empregada para a segunda forma, e pode se beneficiar muito do alinhamento de ontologias (*ontology matching*, no inglês).

Trabalhos importantes sobre extração de informações em patentes podem ser encontrados em Ghoula et al. [1], que descreveu um método para gerar anotações semânticas em textos de patente, usando a estrutura do documento e um esquema de anotação multinível em ontologias, auxiliado por uma combinação de técnicas PLN. Apesar desta abordagem ser veloz e bem alinhada com uma perspectiva de web semântica, ela depende de documentos estruturados e da existência de uma ontologia de domínio para a extração de informação dentro das reivindicações. Taduri et al. [2] propôs uma ontologia para sistemas de patentes, objetivando padronizar a representação a partir de diferentes fontes de informação, inicialmente com foco nos registros do escritório de patentes e cortes judiciais dos EUA. Yang e Soo [3] apresentaram um método para extração de grafos conceituais a partir de reivindicações, usando informações sintáticas e uma ontologia de base, também focando na estrutura das patentes dos EUA.

Bach et al. [4] propôs um método para reconhecimento de partes lógicas e estruturas lógicas previamente definidas a partir de parágrafos em artigos jurídicos japoneses, usando uma combinação de métodos estatísticos e de Programação Linear Inteira. As definições propostas para as partes lógicas e estruturas lógicas, em conjunto com os modelos de aprendizado multicamada e de resolução de restrições em grafos que são usados para segmentação de sentenças e parágrafos, são uma aplicação de modelagem de funções conceituais.

Para a Língua Portuguesa, Ferreira et al. [5] elaborou um método para a extração de relações não taxonômicas entre conceitos, combinando a extração de conceitos usando informação sintática, com uma abordagem estatística centrada em verbos para a extração de relações. Bruckschen et al. [6] apresentou um método baseado em regras para extração de relações entre entidades nomeadas e Caputo [7] uma abordagem de *clustering* para encontrar relações semânticas em patentes brasileiras, usando os campos de sumário e metadados dos documentos.

Trabalhos relacionados sobre análise semântica incluem o método para extração de relações parte-todo independente de domínio, apresentado por Girju et al. [8], e

o algoritmo fracamente supervisionado para extração de relações com padrões genéricos apresentado por Pantel e Pennacchiotti [9], ambos usando aprendizado de anotações semânticas. Nguyen e Shimazu [10] desenvolveram um método semisupervisionado para decomposição semântica, o qual aplica uma técnica de extração de funções conceituais, na forma de representações de significado. Neste, um modelo de alinhamento semântico generativo é utilizado para mapear sentenças às representações de significado para realizar o treinamento.

O trabalho desenvolvido difere dos trabalhos relacionados na forma como é tratado o problema de generalização estrutural das sentenças, usando anotações semânticas para reivindicações de patentes como em [1] e informação sintática como em [3], mas com uma abordagem independente de ontologias. Esta independência foi considerada importante uma vez que os ganhos obtidos com o uso de ontologias ainda poderiam ser aplicados sobre os resultados deste trabalho, podendo melhorá-los. Entretanto, embutir tal recurso apenas aumentaria a complexidade da avaliação do sistema como um todo.

A abordagem adotada neste trabalho trata primeiro da segmentação semântica para a partir desta obter as relações entre os segmentos. Esta ordem é oposta a de [6] e [5], onde as unidades de significado são obtidas a partir das relações que as evidenciam. Esta diferença favorece o uso deste trabalho em uma variedade maior de tipos de texto, embora aumente a dependência do conjunto de dados de treinamento. Além disso, o foco deste trabalho é a obtenção de informação detalhada a partir das reivindicações, a partir da qual a sumarização dos documentos torna-se uma tarefa de composição das informações obtidas.

No aspecto da análise semântica, este trabalho é independente de domínio como [8], através da modelagem e extração de funções conceituais como em [9] e [10], mas propondo um conjunto de atributos diferente e um método de alinhamento novo, utilizando a árvore de constituintes sintáticos. Apesar da necessidade de supervisão, este alinhamento permite classificar com precisão estruturas mais complexas quando comparado com [9] e [10], algo essencial na análise de reivindicações de patentes.

1.3 Objetivos

Os objetivos gerais do trabalho foram:

1. Entender as unidades de significado de uma reivindicação de patente e elaborar um método para identificação destas, a partir de exemplos.

2. Desenvolver um método para capturar as relações entre as unidades de significado identificadas, de forma a reconstruir a informação da patente de maneira estruturada.

3. Construir um sistema de extração de informações a partir de (1) e (2).

Os objetivos gerais foram traduzidos em objetivos específicos, listados abaixo:

1. Obter um conjunto de características que permitissem distinguir entre partes diferentes de uma sentença (segmentos), de acordo com seu significado no discurso;

2. Elaborar um esquema de anotações para expressar de forma simples os segmentos em uma sentença, para ser usado por pessoas, e de fácil tratamento computacional;

3. Desenvolver um método para unir as anotações de segmentos feitas por pessoas usando o esquema (2) àquelas feitas por ferramentas de Processamento de Linguagem Natural, em especial analisadores sintáticos, para encontrar associações relevantes entre ambas.

4. Desenvolver um sistema baseado em aprendizado de máquina para coletar exemplos de reivindicações anotadas usando o método (3) e produzir em reivindicações inéditas anotações consistentes com os padrões aprendidos.

5. Desenvolver um sistema para extração de relações entre os segmentos anotados pelo sistema (4).

Dada a abrangência do tópico de extração de informações em patentes, o escopo deste trabalho foi limitado à análise das informações superficiais contidas apenas nas reivindicações de patente. Desta forma, este trabalho não se propõe a fazer uma análise do conteúdo integral do documento de patente, não sendo uma alternativa a sistemas de indexação de documentos. Além disso, o método desenvolvido visa contornar dificuldades específicas da análise textual de patentes através da generalização sobre estruturas linguísticas, sem o uso de heurísticas específicas para patentes. Logo não objetiva competir com sistemas especialistas de extração de informação de patentes.

1.4 Estrutura da dissertação

O Capítulo 2 apresenta os conceitos que são utilizados ao longo do trabalho e como se relacionam.
No Capítulo 3 o sistema desenvolvido que implementa os métodos elaborados no trabalho é descrito em detalhes, especificando as entradas e saídas, o modelo de processamento utilizado e seus princípios de funcionamento.
No Capítulo 4 são descritos os aspectos considerados para a avaliação do funcionamento do sistema e o procedimento experimental, bem como os resultados obtidos

nos testes aplicados ao sistema.
O Capitulo 5 conclui a visão do trabalho, comentando os resultados alcançados. São apresentadas sugestões de melhorias e expansões para o trabalho.

Capítulo 2

Conceitos básicos

2.1 Patentes e proteção à propriedade industrial

Uma patente é um título emitido pelo governo, que concede propriedade temporária sobre uma invenção. A patente é outorgada pelo Estado aos inventores, autores ou outras pessoas físicas ou jurídicas detentoras de direitos sobre a criação. Aquele que recebe o direito de propriedade deve descrever detalhadamente o conteúdo da invenção, visto que o direito será concedido apenas sobre aquilo que foi descrito. O conteúdo detalhado pelo requisitante da patente é chamado de *escopo* da patente, pois delimita aquilo que pode ou não ser considerado parte da invenção.

As patentes são requisitadas por indivíduos ou organizações que desejam obter proteção legal sobre a exploração de seus inventos, garantindo o direito aos benefícios obtidos pela aplicação dos mesmos. Para fazer uso comercial de um invento patenteado, deve-se obter autorização do detentor da patente, que geralmente cobra uma taxa para isso. Desta forma, as patentes viabilizam a geração de renda através da atividade criativa e com isso incentivam o investimento em inovação.
Tipicamente, empresas já requisitam patentes como uma forma de proteger suas tecnologias de concorrentes. Indivíduos e universidades requisitam patentes para obter retorno pelo esforço de invenção e também o reconhecimento de sua relevância pela sociedade.

Entretanto, se os benefícios da concessão de patentes são importantes, problemas decorrentes da aplicação do direito de propriedade também são de grande relevância. Um indivíduo ou organização que julgue ter tido sua patente infringida deve provar que o alegado infrator está fazendo uso não autorizado de sua invenção. Este por sua vez tentará provar que a invenção em disputa não está coberta pela patente invocada na acusação. Desta forma, para evitar infrações de patente e custosas disputas judiciais, empresas muitas vezes precisam fazer uma pesquisa de patentes para saber se o produto que desejam comercializar não está fazendo uso de alguma invenção já

patenteada. O mesmo tipo de pesquisa pode ser feita para saber se algum produto concorrente está infringindo uma patente já obtida. O custo de requisição e pesquisa de patentes não é baixo e tende a crescer com a quantidade e complexidade dos produtos produzidos. Isto acaba criando uma vantagem para organizações maiores, que dispõem de mais recursos financeiros e podem pesquisar e obter uma grande quantidade de patentes e utilizá-las para limitar as opções de inovação de concorrentes menores. Patentes exageradamente abrangentes também causam problemas ao permitir caracterizar invenções alheias dentro de seu escopo. Todos estes problemas devem ser avaliados e resolvidos durante o processo de concessão ou durante uma disputa judicial sobre uma patente, sendo esta última opção uma possível evidência de que houve falha na concessão, podendo a patente já concedida ser invalidada.

O processo de concessão de patentes é feito pelo escritório de patentes de cada país. No caso do Brasil o *INPI* (Instituto Nacional de Propriedade Industrial) desempenha esta função. Cabe ao escritório de patentes analisar os pedidos, verificando seu grau de similaridade com invenções já patenteadas e se a invenção descrita é patenteável. Há restrições sobre o que pode ser patenteado, baseado no tipo e escopo da invenção. Exemplos notáveis de invenções que não podem ser patenteadas são as fórmulas matemáticas, substâncias naturais e programas de computador.

2.1.1 Etapas da concessão de uma patente

O processo de obtenção de uma patente varia conforme o país onde esta é requisitada. Algumas etapas comuns são apresentadas a seguir:

Busca prévia Consiste na pesquisa por patentes de conteúdo similar à daquela que se pretende requisitar, nos arquivos do escritório de patentes. Esta etapa não é obrigatória, mas é geralmente feita para poupar tempo e despesas burocráticas desnecessárias caso o pedido de patente venha a ser rejeitado em uma etapa posterior por já existir patente muito similar. No INPI, a busca pode ser feita pelo requisitante ou por um técnico do escritório de patentes, com taxas para ambas as formas, sendo a última a mais cara.

Depósito do pedido de patente Preenchimento do *documento de pedido de patente* pelo requisitante e sua entrega ao escritório de patentes. No INPI é cobrada uma taxa para o depósito do pedido de patente.

Exame formal preliminar O documento de pedido é analisado para verificar se foi corretamente preenchido e se seu conteúdo está de acordo com exigências do escritório de patentes. Documentos de pedido que não atendem às exigências são devolvidos com um prazo para correção, depois do qual são arquivados.

Publicação O pedido de patente é divulgado publicamente, para que partes interessadas possam se manifestar a respeito da patente, e possivelmente tentar contestá-la. O INPI publica os pedidos de patente dezoito meses após o depósito do pedido, sendo disponibilizados apenas os dados identificadores do pedido, o resumo e uma figura. A publicação pode ser antecipada pelo INPI mediante o pagamento de uma taxa.

Solicitação do exame do pedido Nesta etapa, o pedido será analisado para verificar se este é de fato uma invenção original e se é patenteável. Além disso, se houve contestações, estas serão também analisadas para apurar seu mérito. No INPI, o exame do pedido só pode ser feito, no mínimo, 60 dias depois da sua publicação.

Exame técnico Durante o exame técnico do pedido, a patente pode ser declarada parcial ou totalmente nula, dependendo do mérito de cada reivindicação. Caso apenas algumas das reivindicações tenham sido anuladas, o requisitante é intimado a manifestar-se, enviando uma versão revisada do pedido. Esta etapa pode se repetir até que o requisitante envie um pedido plenamente aceito ou desista do pedido, que será arquivado.

Expedição da Carta-Patente Uma vez aprovado no exame, será emitida a Carta-Patente, que é o documento oficial comprovando a titularidade da patente.

Manutenção A partir da publicação do pedido de patente, o requisitante deve pagar uma taxa anual para manutenção da patente. Esta taxa deve ser paga enquanto a patente estiver em vigor e aumenta de acordo com a idade da patente.

2.1.2 Estrutura do documento de pedido de patente

Preâmbulo Contém as informações básicas à respeito da patente: Título, identificador, autor(es) e datas, bem como um resumo do conteúdo da patente, descrevendo as características principais do invento.

Relatório descritivo Descrição detalhada do invento, opcionalmente ilustrada.

Reivindicações Uma reivindicação descreve o conceito, processo ou material que é o objeto da proteção legal em uma linguagem estruturada e mais precisa que o restante do documento de pedido de patente.

Ilustrações Desenhos técnicos ou esquemas, que ilustram visualmente os elementos apresentados nas reivindicações e no relatório descritivo.

Resumo Descrição resumida do invento, contendo todas as características principais.

Tipos de reivindicações

As reivindicações podem ser classificadas em dois tipos, dependendo de sua função relativa a um objeto de proteção legal da patente: reivindicação independente e reivindicação dependente.

Reivindicação independente: declara um objeto de proteção da patente, indicando suas características básicas. Exemplo:

"*Blindagem protetora contra arrombamento de cofres compreendendo um conjunto de painéis de blindagem (31, 32, 33, 34, 35) caracterizada pelo fato de ditos painéis estarem instalados no interior do cofre, mediante justaposição às faces internas das paredes, piso e teto do cofre.*"

Reivindicação dependente: detalha um objeto previamente declarado, referenciando uma ou mais reivindicações envolvidas no detalhamento. Exemplo:

"*Blindagem protetora de acordo com a reivindicação 1, caracterizada pelo fato de ditos painéis estarem solidamente unidos entre si ao longo de suas bordas.*"

Grafo de reivindicações

As relações de dependência entre reivindicações podem ser representadas na forma de um grafo direcionado. Em um documento de patentes bem construído, este grafo é acíclico e toma a forma de uma árvore, chamada de *árvore de reivindicações.*
Uma árvore de reivindicações é construída tomando como raiz a reivindicação independente e ligando a ela as reivindicações dependentes que a referenciam. Cada uma destas é ligada às suas reivindicações dependentes e assim por diante. Patentes com mais de uma reivindicação independente possuirão múltiplas árvores e patentes cujas reivindicações dependentes possuam mais de uma referência não formarão árvores, mas sim grafos direcionados acíclicos comuns. A Figura 2.1 ilustra um grafo de reivindicações.

O grafo de reivindicações é uma estrutura importante para a compreensão de como a patente está constituída e como seus componentes se relacionam. Durante a etapa de exame técnico do processo de concessão, reivindicações podem ser anuladas e o efeito das anulações pode ser visualizado com maior facilidade através do grafo, simplificando possíveis correções.

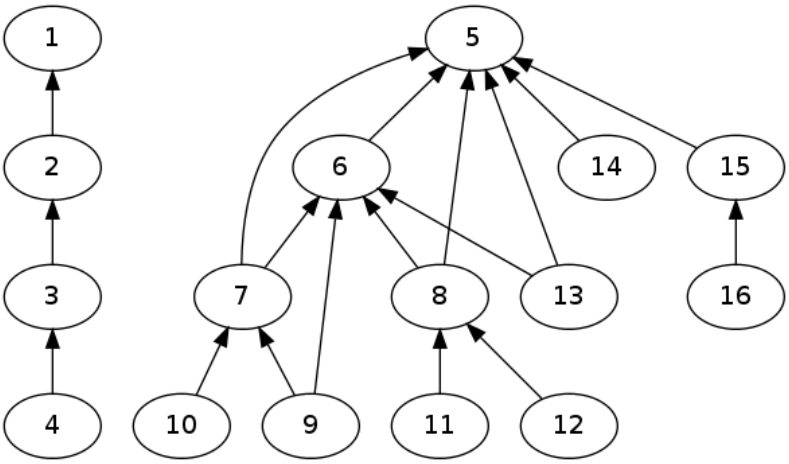


Figura 2.1: Árvores de reivindicações da patente PI0803602-0A2 "Blindagem Protetora Contra Arrombamento de Cofres" (Apêndice C). As reivindicações 1 e 5 são independentes e as demais dependem dos nós adjacentes na árvore. As arestas apontam na direção da dependência. Reivindicações podem possuir múltiplas dependências, se assim especificado no texto da reivindicação.

2.1.3 Outros documentos

Dependendo do escritório de patentes, durante o processo de concessão podem ser gerados vários outros documentos além do pedido de patente. Dentre eles podem ser citados os documentos de exigência (separação de pedidos, exclusão de reivindicações, etc), gerados quando todo ou parte do pedido é reprovada no exame técnico, e os pedidos de contestação. Estes documentos contém informações sobre a patenteabilidade de certas reivindicações, sendo uma importante fonte de metadados para patentes.

2.2 Representação do conhecimento: Ontologias para patentes

Representar o conhecimento de patentes não é uma tarefa trivial. A abrangência de tópicos cobertos pelas invenções faz com que muitos termos tenham significados completamente distintos em documentos diferentes. Termos similares em uma área de conhecimento podem ser diferentes em outra, tornando a pesquisa por palavras pouco útil. Neste contexto, o uso de ontologias apresenta-se como uma solução robusta, e com isso amplamente aceita para a representação deste tipo de conhecimento.

No contexto das Ciências da Computação e da Informação, *ontologia* é uma especificação de um conjunto de primitivas usado para modelar formalmente o conhecimento sobre uma realidade, seja ela física ou virtual. Estas primitivas são

tipicamente *classes*, *atributos* e *relações*, que representam respectivamente os conceitos conhecidos, as características que podem detalhá-los e as formas como se relacionam [11]. Uma ontologia tipicamente descreve também instâncias dos conceitos nela contidos (indivíduos ou exemplares), sendo neste caso também chamada de *base de conhecimento* (Seção 2.2.4) em sua manifestação física, e.g., banco de dados ou arquivo de registros. Como uma especificação formal, a ontologia tem o objetivo de não apenas representar o conhecimento, mas também permitir inferências sobre os objetos nela contidos ou dela instanciados. Os indivíduos, conceitos, características e relações têm seus significados codificados na ontologia, e estes significados podem ser mapeados nos termos utilizados em um texto. A Figura 2.2 ilustra o mapeamento entre os termos de um texto, a ontologia e o significado codificado nesta.

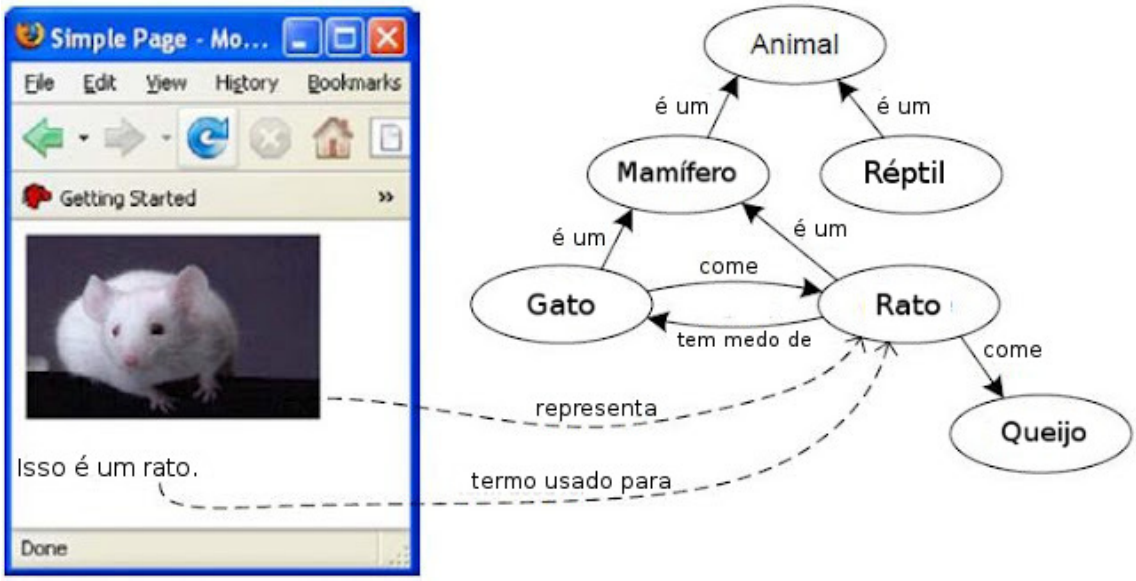


Figura 2.2: Exemplo de mapeamento de elementos textuais e extratextuais em uma ontologia. A palavra "rato" e a figura apresentam ligações de tipos diferentes ao conceito "Rato" do fragmento de ontologia exibido.

No caso dos documentos de patente, o mapeamento de termos a uma ontologia serve para delimitar o domínio de conhecimento (contexto) usado em cada documento, resolvendo o problema de ambiguidade resultante da grande abrangência de tópicos.

2.2.1 Ontologias de domínio

O uso mais típico da palavra ontologia, conforme a definição acima apresentada, refere-se à *ontologia de domínio*. Uma ontologia de domínio tem como objetivo modelar uma área específica do conhecimento, onde há idealmente um único significado para cada termo, permitindo a construção de conceitos sem ambiguidades. Por

exemplo: o termo "carvão" certamente possuirá significados distintos no domínio de mineração e no de culinária. Utilizando a ontologia de domínio apropriada, o termo poderá ser mapeado ao conceito adequado na ontologia, i.e., carvão mineral e carvão vegetal, respectivamente, e o significado correto será obtido.

Exemplos de ontologia de domínio incluem: *Foundational Model of Anatomy* [12] (anatomia humana), *Disease Ontology* [13] (patologia humana), *FAO* [14] (geopolítica) e *FOAF* [15] (pessoas).

2.2.2 Ontologias superiores

Enquanto as ontologias de domínio modelam conhecimento específico, há conceitos que são generalizáveis para todos os domínios. Para modelá-los são utilizadas as *ontologias superiores ou gerais*, que têm como função permitir a interoperabilidade entre ontologias. A generalização de conceitos é feita de acordo com a finalidade dada à ontologia e por este motivo existe uma variedade de ontologias superiores, tais como: *Suggested Upper Merged Ontology* (SUMO) [16] e *Basic Formal Ontology* (BFO) [17].

2.2.3 OWL & RDF

Para que uma ontologia saia do campo teórico e possa ser utilizada em computadores, faz-se necessária sua representação em uma *linguagem formal*. Ao longo do tempo, diferentes linguagens foram propostas para a representação de ontologias, variando em expressividade conforme o tipo de informação a ser armazenado e o tipo de processamento desejado sobre as informações, tais como buscas e inferências.

As duas linguagens mais utilizadas atualmente para a representação de ontologias são:

OWL [1] [18]: Compreende uma família de linguagens baseadas em *XML* [19], projetadas para distribuição na web, e tendo a *Lógica de Descrição* [20] como modelo semântico formal. A grande expressividade que pode ser obtida com OWL tornou esta a recomendação oficial para publicação de ontologias na web.

RDF [2] [21]: Linguagem para modelagem conceitual, baseada em "*afirmações*" ou "*fatos*" apresentados sobre *recursos*, na forma de expressões sujeito-predicado-objeto chamadas de *triplas*. Um recurso em RDF pode ser qualquer elemento real ou virtual descrito em uma *URI* [3] ou um dado primitivo, como um número, uma data ou um

[1] Web Ontology Language
[2] Resource Description Framework
[3] Uniform Resource Identifier

caractere. Em uma tripla RDF, o sujeito e objeto são recursos, enquanto o predicado descreve a relação, e.g., propriedade, ação, ligando ambos.

Fatos descritos em algumas linguagens OWL podem ser diretamente mapeados em RDF correspondente. Um conjunto de fatos RDF representa um multigrafo direcionado onde cada nó representa um recurso e cada aresta, um predicado. Tal modelo de representação é também adequado a fatos obtidos a partir de sentenças declarativas, facilitando a instanciação de conceitos em ontologias.

2.2.4 Instanciação de conceitos: bases de conhecimento

A descrição de um conceito em uma ontologia pressupõe a existência de pelo menos um exemplar (instância) que possa ser identificado como pertencente à classe que representa tal conceito. A atribuição de um exemplar a uma classe, com o preenchimento dos valores correspondentes às características descritas para a classe, bem como suas relações com outros objetos, é chamada de *instanciação de conceito* ou simplesmente *instanciação*. Exemplo de instanciação: a Terra pode ser considerada um exemplar da classe "Planeta" descrita em uma ontologia e desta forma ter suas características, e.g., período de rotação e gravidade superficial, e suas relações, e.g., distância de outros planetas, preenchidas com seus dados. Estes dados podem ser fornecidos por um ser humano ou automaticamente, através de consultas a fontes de dados estruturadas ou semiestruturadas, como um banco de dados, ou de processamento de texto em linguagem natural.

As instâncias podem ser representadas na mesma linguagem usada para o restante da ontologia. Quando ambos o modelo conceitual e as instâncias de uma ontologia podem ser acessados em um mesmo local, este conjunto (modelo + instâncias) é chamado de *base de conhecimento*. Como exemplo de base de conhecimento, a DBpedia [22] preenche suas instâncias com dados semiestruturados obtidos da popular enciclopédia colaborativa Wikipédia [4].

2.2.5 Wordnets

Uma *wordnet* é uma base de dados léxica, que combina as funções de um *dicionário* e um *tesauro*[5], tendo seu nome e origem na *WordNet* [23] da Universidade de Princeton (EUA). Ela agrupa as palavras de uma determinada língua em conjuntos de sinônimos chamados *synsets*, cada um contendo uma breve explicação sobre seu uso e contexto e também as relações semânticas com outros synsets. As *wordnets* servem pricipalmente para a desambiguação de significado das palavras, e com isso

[4] http://www.wikipedia.org/

[5] Lista de palavras agrupadas por similaridade de significados, mostrando apenas as diferenças, mas sem as definições

oferecem uma grande ajuda para tarefas de classificação e sumarização automatizada de textos. O crescimento do número de línguas contempladas por *wordnets* contribui também para avanços na tradução automática de textos.

As *wordnets* apresentam muitas semelhanças à ontologias, pois também representam conceitos (os synsets) e suas relações, mas diferem destas pois não possuem uma especificação formal e, por isso, podem conter inconsistências. São, entretanto, muito menos complexas, e tal semelhança permite o uso de *wordnets* no lugar de ontologias em situações em que o formalismo não seja estritamente necessário, em troca de resultados mais próximos da linguagem humana e menor custo computacional.

2.3 Análise semântica de texto em linguagem natural

Em linguística, *Semântica* é o estudo do significado das palavras, sinais, símbolos, frases ou expressões em um determinado contexto. Também trata das relações entre os significantes e o que estas relações representam. Tal definição é também empregada no campo de estudo do *Processamento de Linguagem Natural*, que busca um tratamento computacional para a compreensão da linguagem humana, e, portanto,, tem na resolução do significado um de seus pilares fundamentais.
Como qualquer processo computacional, a análise de texto em linguagem natural é realizada em etapas discretas. Estas etapas são correspondentes aos níveis de composição do texto, do mais simples ao mais complexo, e aos diferentes tipos de informação contidas em cada nível. Elas são descritas ao longo desta seção.

2.3.1 Decomposição do texto

Tokenização

Trata da identificação dos elementos mínimos que compõem o texto em uma determinada língua, para sua separação. Em idiomas ocidentais isto é geralmente feito para cada palavra e pontuação, que são típicamente separados por espaços. Dependendo do nível de detalhe requerido, o esquema de tokenização pode levar em conta palavras que são composições de outras e separá-las, como no caso "*da*" = "*de* + *a*". A tokenização é a etapa fundamental da decomposição do texto, pois todas as outras baseiam-se na premissa de que as palavras de uma língua são conhecidas e sempre separadas de uma determinada maneira.

Análise morfológica

Trata da classificação das palavras conforme sua estrutura de formação, olhando para elas isoladamente, em chamadas *classes gramaticais*, *classes morfológicas* ou *Part-of-Speech (POS) tags*: *verbo*, *substantivo*, *adjetivo*, *artigo*, entre outros. Isto pode ser feito sem que seja considerado seu significado em uma sentença, através da análise das palavras vizinhas. As classes gramaticais variam conforme a língua. Em função disto, diversos estudos foram feitos para a obtenção de conjuntos de classes comuns a várias línguas, como em [24] [25], e também para a utilização de contexto gramatical, e.g., palavras vizinhas, na determinação da classe [26] [27].

As técnicas usadas para análise morfológica incluem modelos probabilísticos [28], heurísticas [29], e sistemas de aprendizado de máquina [30] [31], entre outros. Abordagens recentes conseguem superar 97% de precisão na classificação gramatical por palavra para algumas línguas, fazendo com que esta seja considerada uma tarefa já bem resolvida no campo de Processamento de Linguagem Natural, ainda que existam espaços para melhorias [32].

Análise sintática

Trata da decomposição dos padrões estruturais da língua, determinados pelas relações entre as palavras e entre as frases que constituem uma sentença, ou seja, como as palavras estão dispostas nas frases e como estas compõe o discurso. A análise sintática pode ser dividida em dois aspectos de interesse: estruturas de constituintes e dependências gramaticais.

Estrutura de constituintes A formação das sentenças (frases, orações, períodos) se dá por meio da composição de palavras em grupos, chamados sintagmas, que são portanto considerados os constituintes da sentença [33] [34]. Esta composição ocorre de forma hierárquica, com as palavras sendo o nível mais baixo e compondo os sintagmas, que se combinam formando frases nominais ou orações, dependendo do seu tipo. As orações por sua vez se combinam formando períodos, sendo estes coordenados ou subordinados dependendo de suas relações de dependência (ver abaixo).

Dependências gramaticais São relações de dependência existentes entre duas palavras, ou entre duas orações, como a dependência existente entre um adjetivo e o nome por ele modificado (dependência do tipo adjunto adnominal).

Ambos os aspectos carregam informações necessárias para a interpretação correta de uma sentença. Estas informações são mais explícitas no caso das dependências gramaticais, já que estas estão fortemente ligadas às dependências entre o significado

das palavras. Entretanto, as estruturas de constituintes contêm uma quantidade maior de informação, pois representam todos os níveis de agrupamento das palavras na sentença.

Para obter uma representação computável desses aspectos, ou seja, explorável utilizando-se de teorias e ferramentas computacionais, faz-se necessária a transformação do texto em uma *linguagem formal*. Isto é feito através da obtenção de uma *gramática formal* que contemple uma porção da língua suficiente para cobrir a maior parte as construções utilizadas nos textos que serão analisados, permitindo interpretá-los do ponto de vista sintático. O tipo de gramática formal mais utilizado para modelar ambos os aspectos é a Gramática Livre de Contexto Probabilística (PCFG) [35], produzindo as chamadas *gramáticas de constituintes* e *gramáticas de dependência* respectivamente. A corretude de uma PCFG está fortemente ligada à qualidade das anotações que definem os agrupamentos sintáticos, ou seja, das classes morfológicas. Desta forma, uma boa análise morfológica é essencial para o sucesso da análise sintática.

A representação resultante da gramática formal para estruturas de constituintes é a *árvore sintática* ou *árvore de constituintes* e para dependências gramaticais é o *grafo de dependências*. Em uma árvore sintática, um nó raiz "S" delimita a sentença, os sintagmas são marcados pelos nós não terminais, e as palavras são as folhas, podendo estas ser marcadas por suas respectivas classificações morfológicas. A Figura 2.3 ilustra uma árvore sintática. Em um grafo de dependências, as palavras são ligadas por arestas direcionadas representando as *funções de dependência gramatical*: determinante, sujeito, entre outros, sendo que as arestas são orientadas do modificador para o modificado. A Figura 2.4 ilustra um grafo de dependências gramaticais. O resultado esperado da análise sintática é portanto a construção de árvores sintáticas ou grafos de dependência, dependendo do tipo de informação desejada.

Análise semântica

Trata da obtenção do significado das palavras ou expressões, tanto isoladamente quanto em frases, e também das relações entre significados. Assim como na análise sintática, a classificação de significados e relações é dependente do seu propósito. Dentre as formas mais populares de classificação destacam-se: os papéis semânticos e as relações semânticas.

Papéis semânticos expressam a função de uma ou mais palavras em relação a um verbo. Por exemplo na frase: "*João comeu a manga.*", João é o agente do verbo comer e portanto o agente da sentença, e da mesma maneira a manga é o termo paciente da sentença.

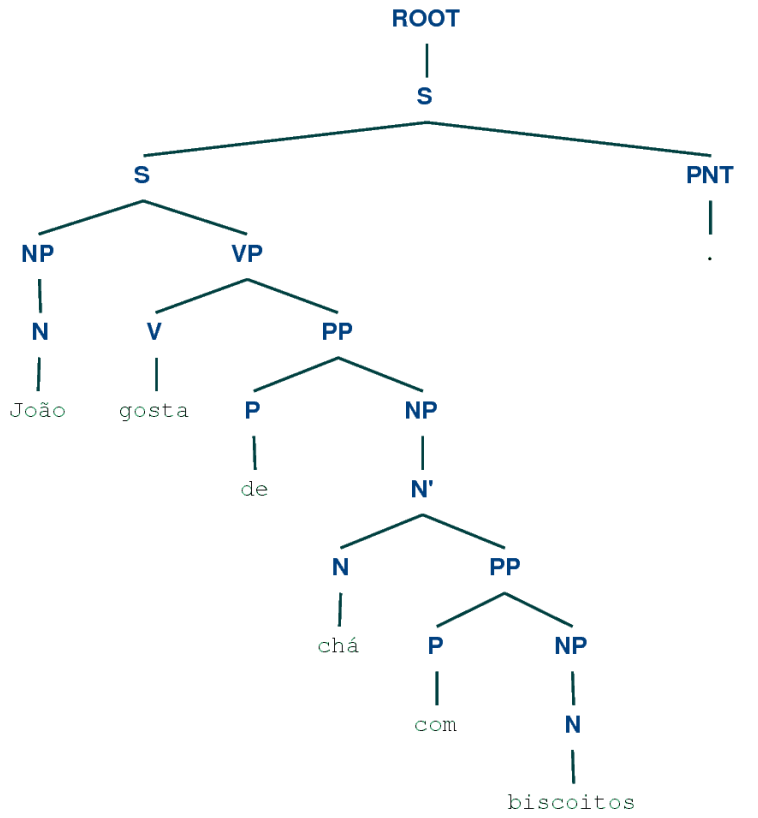


Figura 2.3: Exemplo de árvore sintática.

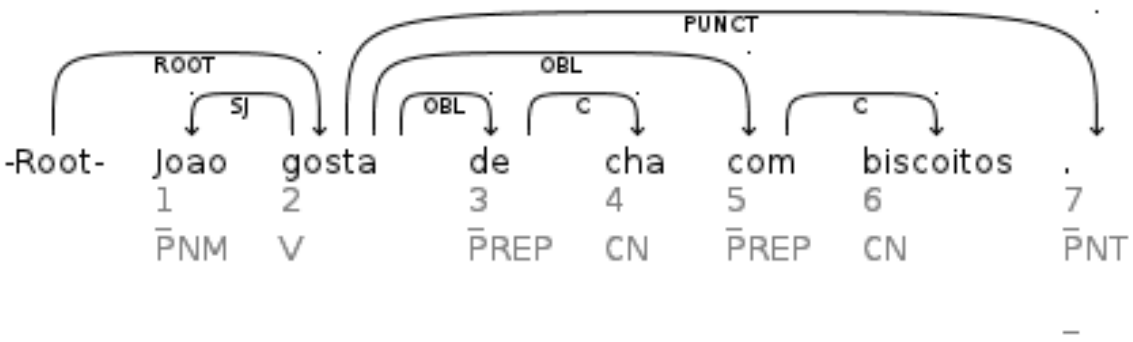


Figura 2.4: Exemplo de grafo de dependências gramaticais.

Relações semânticas expressam a relação existente entre o significado de duas palavras ou expressões. Podem ser divididas em 3 tipos principais: *ação ou estado*, e.g., "trabalha em"; *característica*, e.g., "adequado a" e *conceitual*, e.g., "sinônimo". Podem ocorrer de 3 modos:

- *Entre conceitos*, e.g., cachorro e mamífero apresentam a relação "*é um*" (conceitual) do primeiro para o segundo termo.
- *Entre conceito e indivíduo*, e.g., Zé Pequeno e policial apresenta, a relação "*antagoniza*" (ação ou estado) para ambos os lados.

- *Entre indivíduos*, e.g., Dilma e Brasil apresentam a relação "*presidente em exercício*" (característica) do primeiro para o segundo termo.

A classificação em papéis semânticos tem origem no trabalho de Gildea e Jurafsky [36] em uma aplicação prática da teoria de *Frame semantics* [37]. A noção estruturada de relação semântica tem origem em trabalhos linguísticos de Lyons [38] e Cruse [39]. Estas formas de classificação passaram por evoluções, conforme ocorreu com suas teorias subjacentes, e ao mesmo tempo em que se tornaram mais abrangentes, também incorporaram formalismos relativos à diversas áreas do conhecimento humano [40]. Outras formas de classificação podem ser encontradas nas teorias *Frame semantics* [37] e *Discourse Representation Theory* [41]. O trabalho de Bean [40] apresenta a noção de relação semântica utilizada neste trabalho: "uma associação entre duas ou mais entidades ou entre duas ou mais classes de entidades".

Assim como na análise sintática, faz-se necessária uma representação formal para o uso computacional de classes semânticas. Neste trabalho são utilizados os conceitos de *Função Conceitual* e *Modelo de Relacionamentos*.

A Função Conceitual [42] é uma abstração lógica para uma unidade de significado no texto, que pode ser constituída de uma ou mais palavras. É representada na forma de um predicado $F(X, \dots)$, onde F é a função e X é um termo participante da unidade de significado denotada pela função. Exemplos: Coisa(o cachorro); Causa(enchente, chuva).

O Modelo de Relacionamentos [43] é uma abstração relacional para a ligação entre um conjunto de unidades de significado, caracterizado pelo uso de grafos direcionados onde os vértices tipicamente denotam conceitos ou indivíduos e as arestas denotam as relações entre estes. O modelo de relacionamentos usado neste trabalho é o de triplas, onde cada relacionamento é mapeado em uma tupla (sujeito, predicado, objeto), que representa o enunciado lógico *predicado(sujeito, objeto)*.

Chomsky, em seu famoso trabalho "Syntactic Structures" [34] sugere a noção de "significado estrutural", onde há pontos importantes de correlação entre as estruturas sintáticas e seus significados. Katz e Fodor [44] desenvolvem esta noção, apresentando o conceito de *regras de projeção*: mapeamentos entre constituintes sintáticos e seus significados, na forma de marcadores semânticos aplicados sobre elementos gramaticais. Jackendoff [42] expande tais ideias e faz a ligação entre as regras de projeção e as funções conceituais na forma de *constituintes conceituais*.

O mapeamento de constituintes sintáticos em conceituais propicia o mapeamento do texto em ontologias, visto que há uma relação de um para um entre constituintes conceituais e os elementos de uma ontologia (conceitos, relações, características, indivíduos), onde cada constituinte está vinculado à uma função conceitual, que abstrai uma unidade de significado, unidade esta representada na ontologia. En-

tretanto, tal mapeamento apresenta uma série de desafios, que serão explorados ao longo deste trabalho.

2.3.2 Segmentação semântica

A classificação do texto em papéis semânticos permite obter uma grande variedade de funções conceituais, sobretudo as mais frequentes no uso típico da língua. Moldovan et al. [45] identifica diversas classes de papéis, cobrindo a maior parte da semântica cotidiana. Entretanto, funções conceituais em tipos diferentes de texto podem ser melhor modeladas por tipos específicos de classes. Este é o caso das reivindicações de patente, que podem cobrir um vasto número de áreas do conhecimento, mas que possuem um conjunto de funções comuns, como "assunto da patente", "referência à reivindicação" e "caracterização de objeto".

Para obter compatibilidade com qualquer conjunto de classes de função conceitual, abstraindo o uso de ontologias, este trabalho utiliza o conceito de *segmento semântico*: uma subsequência qualquer de palavras em uma sentença, para a qual uma função conceitual pode ser atribuída [46]. O segmento semântico é uma generalização do conceito de papel semântico, onde a classe do segmento representa uma função conceitual relativa a qualquer elemento dentro ou fora da sentença. Por exemplo, a classe *NUM_REF_REIVIND* indica o número usado para referenciar uma reivindicação específica no documento de patente. A Figura 2.5 ilustra uma sentença segmentada semanticamente para a reivindicação de patente *"Blindagem protetora de acordo com a reivindicação 8, caracterizada pelo fato de ditas substâncias inorgânicas compreenderem o cloreto de amônio."*. Nesta sentença, podem ser evidenciados os seguintes segmentos, relativos ao contexto de uma patente:

- O assunto (tópico) da patente: "Blindagem protetora".
- A referência a um outro elemento do texto, nesse caso a "reivindicação 8".
- O número identificador da referência: "8".
- A caracterização explícita do assunto: "caracterizada pelo fato de ...".
- Objetos de proteção legal da patente: "substâncias inorgânicas ...".
- Caracterização desses objetos: "compreenderem o cloreto de amônio".

É dado o nome de *Segmentação semântica* à tarefa de identificação e classificação de segmentos semânticos em uma sentença. Esta tarefa é a base utilizada neste trabalho para o mapeamento de constituintes sintáticos em conceituais, através da técnica de *alinhamento sintático-semântico* (Seção 3.5.1). O resultado esperado da segmentação semântica é a identificação correta de todas as funções conceituais em uma sentença e a atribuição dos termos relativos a cada função.

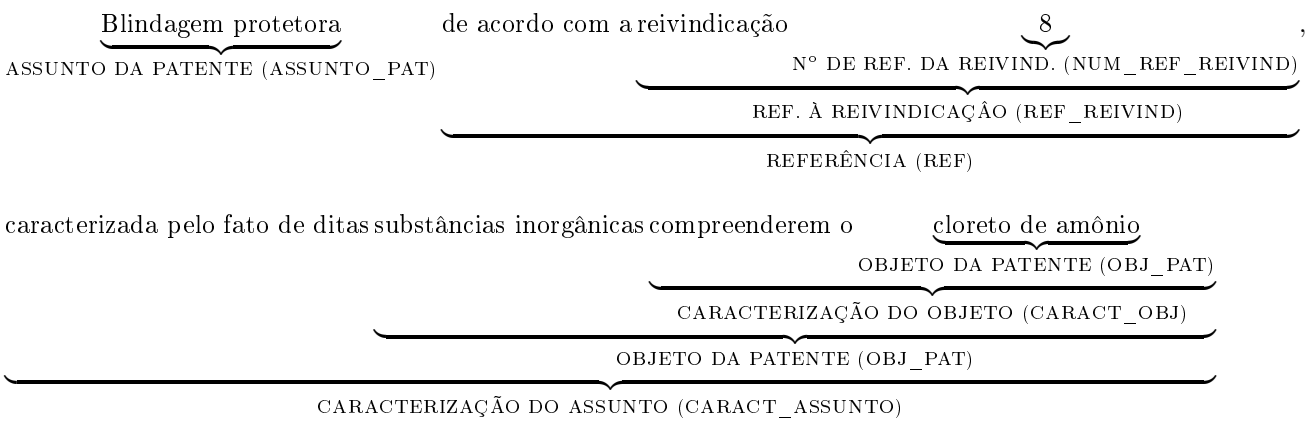


Figura 2.5: Sentença segmentada semanticamente.

2.3.3 Resolução de Entidades Nomeadas

Entidades Nomeadas (Named Entities - NEs) são coisas ou pessoas que podem ser referenciadas por termos, simples ou compostos, de conhecimento público. Por exemplo, os termos "Mona Lisa" e "Gioconda" referem-se a uma mesma obra de arte específica, da mesma forma que "Jorge Mario Bergoglio" e "Papa Francisco" referem-se a uma mesma pessoa. A principal característica de uma entidade nomeada é referir-se a apenas um indivíduo e não a um conjunto. Portanto "carro" e "taxista" não são entidades nomeadas. A literatura de linguística computacional apresenta várias definições para entidades nomeadas, mas a definição mais comumente aceita pode ser encontrada em [47]. As entidades nomeadas consideradas neste trabalho são aquelas para as quais há uma instância registrada em uma ontologia, permitindo seu tratamento computacional.

A tarefa de atribuir um termo à identidade de sua entidade correspondente é chamada de *Resolução de Entidade Nomeada* (conhecida pela sigla em inglês para *Named Entity Recognition* - NER). Ao tratar de documentos de patente, esta tarefa compreende a identificação dos termos de interesse, e.g., tópicos, objetos de proteção legal, e o mapeamento aos seus respectivos significados para o domínio de conhecimento coberto pelo documento.

As abordagens para NER desenvolvidas até o presente momento são predominantemente estatísticas, com novas técnicas fazendo uso de cada vez mais características intra e intertextuais [48] [49]. Tais abordagens são dependentes do domínio de conhecimento, fazendo com que sistemas NER desenvolvidos para um domínio não funcionem bem em outros. Pereira [50] propõe um método para seleção automática das bases de dados usadas para NER, de acordo com propriedades do texto, além de critérios para avaliação da qualidade das bases.

2.3.4 Resolução de correferência

No estudo linguístico do discurso, designa-se *correferência* um conjunto de figuras de linguagem onde um termo é usado para mencionar a outro no mesmo texto, geralmente na forma de pronomes. Por exemplo, na frase "*A capivara seguia João porque ele tinha comida*", o pronome *ele* refere-se ao termo *João*. Na língua portuguesa, as correferências incluem os chamados pronomes *anafóricos*[6] e *catafóricos*[7]. Além destas, há as correferências não pronominais, e.g., "*Dilma pediu um avião mais rápido, pois a presidente precisa reduzir o tempo gasto em viagens*", que também são conhecidas como correferências *não anafóricas*. Uma análise completa sobre correferências para a língua portuguesa pode ser encontrada em [51].

É chamada de *resolução de correferência* a tarefa de encontrar as referências a um mesmo termo no texto. Abordagens atuais para esta tarefa são predominantemente baseadas em heurísticas, opcionalmente complementadas por métodos estatísticos [52] [53] [54].

Em documentos de patente, as correferências ocorrem principalmente na forma de expressões iniciadas pelos pronomes demonstrativos "tal [...]" e "cujo(a) [...]", e também na forma não pronominal "dito [...]", como ilustrado no trecho abaixo:

"*... caracterizada pelo fato de* ***ditos painéis*** *estarem solidamente unidos entre si ...*"

Neste trecho, o termo referido por "ditos painéis" pode estar localizado na mesma sentença ou em uma sentença anterior do texto. Este é o cenário tipico da resolução de correferência, onde a resposta correta permite obter uma representação única dos termos referenciados (*Normalização de correferências*) e com isso agrupar todos os fatos declarados sobre tais termos.

2.3.5 Extração de relações semânticas

Chama-se *Extração de Relações Semânticas* a tarefa de obtenção das relações semânticas encontradas no texto e sua representação em um modelo de relacionamentos, permitindo seu uso por qualquer aplicação que possa fazer uso do modelo. As abordagens para esta tarefa são muito diversas e dependem do domínio de conhecimento dos textos sendo analisados. As mais populares envolvem o uso de ontologias de domínio para obtenção das classes de relacionamento a serem procuradas no texto, assim como seus sujeitos e objetos. Estas podem ou não fazer uso de informações

[6] Pronome que estabelece uma referência dependente com um termo antecedente na frase
[7] Pronome que faz referência a um termo subsequente na frase

morfossintáticas do texto, e.g., aplicando regras de extração em grafos de dependência gramatical [55], ou utilizando os dados ontológicos para obter os termos relacionáveis via NER e as relações através de comparação por dicionário [56]. Para este último caso, o uso de *wordnets* em conjunto com ontologias também vem ganhando popularidade [57]. Para extrações independentes de domínio, as abordagens estatísticas são mais populares. Recentemente, a disponibilidade de enormes bases de conhecimento e corpora [8] na web, tornou viável o uso de algoritmos para extração e validação de relações diretamente de textos na internet, assim como tornou desejável o mapeamento dos modelos de relacionamento em tais recursos.

O produto final da extração de relações semânticas é uma instância do modelo de relacionamentos adotado, e.g., triplas, formando um *grafo de relações semânticas*, cujos nós e arestas podem ser mapeados respectivamente em instâncias (ou conceitos) e relações de uma ontologia já existente, complementando-a. O grafo também pode ser considerado um pequeno bloco para construção de uma nova ontologia.

2.4 Redes Neurais sem Peso e o modelo WiSARD

2.4.1 Redes neurais tradicionais vs Redes Neurais Sem Peso

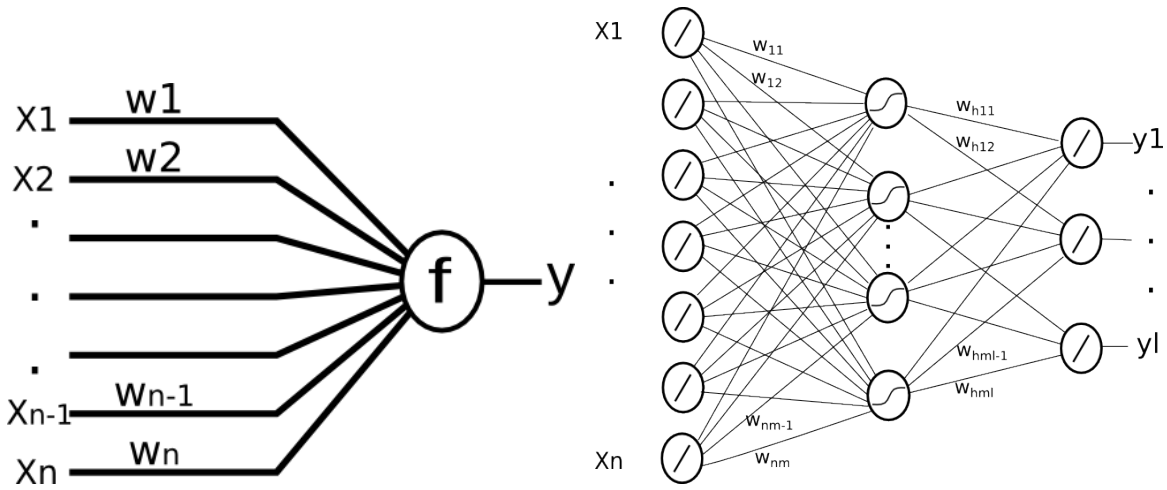
No início da década de 1940, uma importante alternativa ao modelo algorítmico de computação era apresentada pelo trabalho de McCulloch e Pitts [58]. Neste trabalho foi desenvolvido um modelo eletrônico que procurava imitar as conexões entre neurônios do cérebro e suas sinapses, com o objetivo de tornar possível a computação de problemas considerados intratáveis pelos algoritmos até então conhecidos. A capacidade computacional neste modelo era alcançada através de propriedades intrínsecas à rede de neurônios, que tornavam triviais para os seres humanos certas tarefas cuja programação era muito complexa ou até mesmo inviável. O modelo de McCulloch e Pitts obteve grande notoriedade quando foi posto em prática no início da década de 1960, pelo trabalho de Rosenblatt [59]: um sistema capaz de reconhecer imagens, chamado *Perceptron*. O modelo de McCulloch e Pitts e o Perceptron são considerados as bases dos modelos hoje conhecidos como *Redes Neurais Artificiais* (*Artificial Neural Networks - ANN*) "tradicionais" [60] [61].

Uma ANN tradicional é composta por unidades básicas chamadas "neurônios", ligados uns aos outros por múltiplas entradas e saídas, denominadas "sinapses". Cada sinapse possui um peso, que é responsável por modificar a saída de um neurônio que será usada como entrada de outro. Cada neurônio possui uma função de ativação, para a qual valores acima de um certo limiar ativam a saída, sendo tipicamente

[8] Plural de *corpus*: conjunto de textos escritos ou falados de uma língua usados para análise, opcionalmente anotados com informações complementares, como classes gramaticais e estruturas sintáticas.

usada a função sigmóide ou tangente hiperbólica, para permitir a representação de não linearidade pela rede. Os pesos podem atenuar (inibir) ou amplificar (excitar) uma entrada ou saída e devem ser ajustados para que a rede consiga desempenhar a função desejada. O ajuste dos pesos (*treinamento*) é feito através de uma variedade de algoritmos, sendo o *backpropagation* (retro-propagação) o mais popular [61]. Este tipo de rede permite classificar um conjunto de entradas numéricas em um número definido de classes, onde as entradas são tipicamente atributos do problema a ser computado, e.g., intensidades de cor de uma imagem ou frequências de um espectrograma de sinal sonoro.

Este tipo de ANN funciona em camadas, aplicando as entradas aos chamados "neurônios de entrada" que compõe a primeira camada. Estes enviam seus valores para a próxima camada, chamada "camada oculta", através das sinapses. A rede pode possuir nenhuma, uma ou mais camadas ocultas. Após a última camada oculta, se houver, os neurônios são ligados à "camada de saída", onde cada neurônio estará associado a alguma classe do problema sendo analisado. Dependendo do tipo de problema, apenas um deles estará ativo em redes com saída binária, ou apresentará o valor mais alto em redes com saída real. Desta forma, a rede propaga um "sinal" de entrada e o modifica por meio de suas sinapses e funções de ativação até a saída, onde o resultado é obtido. Por este motivo, este tipo de rede é também conhecido como *feedforward*. Uma ilustração do modelo é mostrada na Figura 2.6.



(a) Neurônio artificial (b) Perceptron multicamada

Figura 2.6: Rede Neural tradicional. A figura (a) mostra a unidade básica da rede, o neurônio artificial com suas entradas $X_1 \ldots X_n$, que são modificadas pelos respectivos pesos $w_1 \ldots w_n$. Cada neurônio possui uma função de ativação f que determina o valor y de sua saída. A figura (b) mostra um Perceptron multicamada, sendo a primeira (entrada) e a última (saída) compostas por neurônios com função de ativação linear e a camada intermediária (oculta) composta por neurônios com função de ativação sigmóide. As saídas $y_1 \ldots y_l$ correspondem as classes do problema a ser tratado.

As Redes Neurais tradicionais são uma boa opção para o tratamento computacional de problemas para os quais não há um modelo matemático conhecido ou para aqueles cujos modelos conhecidos são muito custosos. Uma vez treinada, sua operação resume-se a uma sequência de somas e multiplicações, que podem ser feitas rapidamente, mesmo em uma grande quantidade de entradas. Entretanto, possuem limitações importantes. Uma delas é que o treinamento via *backpropagation* prevê a aproximação da solução desejada pelo método de descida do gradiente, o que implica em um número indefinido e possivelmente grande de iterações. Este fato torna o treinamento um processo lento, especialmente com muitas camadas. Além disso, uma rede sem camada oculta está limitada a representar problemas linearmente separáveis, falhando em resolver funções simples como o "ou exclusivo" (XOR) [62].

Uma alternativa ao modelo proposto por McCulloch e Pitts para reconhecimento de padrões é encontrada no modelo de *n-tuplas*, que tem origem no fim da década de 1950, pelo trabalho de Bledsoe e Browning [63]. Neste modelo, uma entrada binária é mapeada em uma matriz, inicialmente preenchida com zeros, selecionando conjuntos aleatórios de k bits (tuplas) que são interpretados como endereços binários para os elementos da matriz. Os elementos endereçados pelos padrões de entrada têm o valor 1 gravado, realizando assim o treinamento do modelo. Ao apresentar um novo padrão como entrada na fase de reconhecimento, pode-se obter uma medida de similaridade, contando quais dos elementos da matriz que foram endereçados estão marcados com 1.

A disponibilidade de memórias eletrônicas permitiu que modelos baseados em n-tuplas fossem implementados em *RAMs* (*Ramdom Access Memory*), na forma de *neurônios-RAM* [64], que passaram a ser conhecidos também como "modelos baseados em RAM". No final da década de 1970, a redução de preço das memórias RAM tornou possível a construção da primeira *ANN Baseada em RAM*, a WiSARD [65], que é descrita na Seção 2.4.2. Diferentemente das ANNs tradicionais, a função desejada nas ANNs baseadas em RAM é ajustada modificando-se o conteúdo armazenado nos neurônios-RAM, em vez de alterarem-se pesos entre sinapses. Por este motivo, também são conhecidas como *Redes Neurais sem Peso* (WANNs - *Weightless Artificial Neural Networks*). Esse fato também implica na ausência da necessidade de convergência do método de descida do gradiente, pois a informação é obtida e armazenada apenas uma vez a cada novo padrão apresentado, tornando o treinamento deste tipo de rede muito mais rápido.

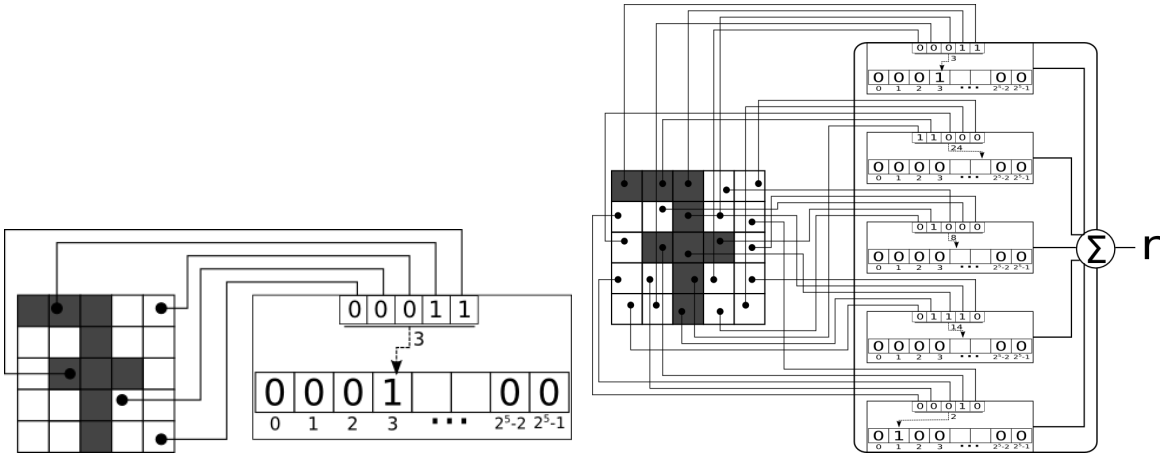
Outros exemplos de WANNs incluem a Memória Esparsa Distribuida (SDM - *Sparse Distributed Memory*) [66], *Goal Seeking Neuron* (GSN) [67], *Generalizing RAM* (G-RAM) [68] e *Virtual G-RAM* (VG-RAM) [69].

2.4.2 O modelo WiSARD

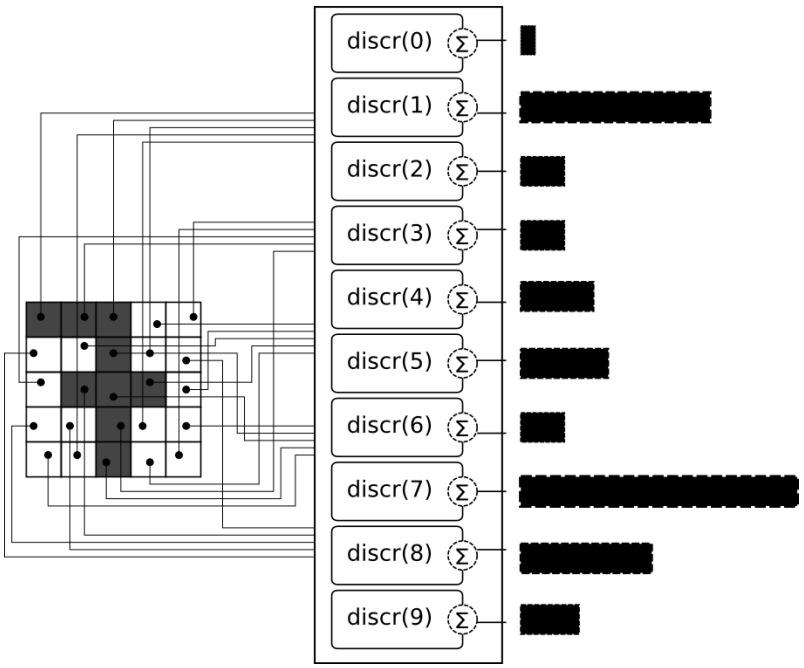
WiSARD (*Wilkie, Stonham & Aleksander's Recognition Device*) [65] é uma WANN formada por vários *discriminadores-RAM*, cada um consistindo de um conjunto de X *neurônios-RAM* com endereços de tamanho n. Um neurônio-RAM consiste em uma memória binária endereçável pela entrada também binária, tendo portanto 2^n posições.

Dessa forma, a rede recebe um padrão binário de $X \times n$ bits como entrada. Em geral, todas as linhas de endereçamento dos neurônios-RAM são conectadas aos bits padrão de entrada por meio de um mapeamento aleatório biunívoco, que permanece constante durante todo o funcionamento da rede. Relembrando, todos os bits dos neurônios-RAM são zerados inicialmente.
O treinamento da rede é feito atribuindo-se "1" às posições de memória endereçadas pelos padrões de entrada (Figura 2.7a. A WiSARD classifica os padrões ainda não vistos somando os conteúdos de memória endereçados e assim obtendo o número de neurônios-RAM que produziram "1" como saída. Este somatório é chamado de *resposta do discriminador* (r) e expressa o grau de similaridade do padrão de entrada com os padrões do conjunto de treinamento. Cada discriminador-RAM é associado a uma classe do problema a ser resolvido, então quando um padrão é dado como entrada, cada discriminador-RAM fornece uma resposta r para este (Figura 2.7b). As respostas de todos os discriminadores-RAM são comparadas e a classe correspondente à maior resposta é selecionada como a classe correta para o padrão de entrada. A Figura 2.7c ilustra a arquitetura do classificador WiSARD.
Um modo clássico de realizar a comparação entre as respostas dos discriminadores é pelo uso da confiança relativa (c) da resposta. Esta pode ser calculada pela fórmula $c = \frac{r_{MAX} - r_{MAX-1}}{r_{MAX}}$, onde r_{MAX} é a maior resposta e r_{MAX-1} é a segunda maior. Este valor indica o grau de certeza da resposta e consequentemente a chance da classe escolhida ser realmente a correta. Se $c = 0$, então há um empate entre as (duas ou mais) maiores respostas, indicando uma ambiguidade da entrada perante as classes correspondentes. A escolha da classe correta neste caso pode ser aleatória ou arbitrária.

Sendo um modelo básico de WANN, a WiSARD é capaz de representar qualquer informação contida nos padrões de entrada, seja de natureza linear ou não, treinando apenas uma vez cada exemplo de entrada. Sua arquitetura simples facilita a implementação da rede em sistemas com recursos computacionais limitados, e.g., sistemas embarcados, *smartphones*. Além disto, o treinamento pode ser intercalado com a classificação (treinamento *online*), tornando este tipo de rede uma escolha eficiente para situações onde a informação a ser modelada pela rede muda com o tempo, como apresentado em [70].



(a) Neurônio-RAM (b) Discriminador-RAM



(c) WiSARD

Figura 2.7: Arquitetura da rede WiSARD. A figura (a) mostra o neurônio-RAM e sua forma de endereçamento. A figura (b) mostra a construção de um discriminador-RAM através da união de um conjunto de neurônios-RAM. A figura (c) mostra o classificador WiSARD completo, com um discriminador para cada classe do problema sendo tratado. Cada discriminador produz uma resposta conforme o grau de similaridade do conteúdo de suas memórias em relação ao padrão de entrada apresentado. Todos os discriminadores recebem a mesma entrada.

Entretanto, a WiSARD também sofre com algumas limitações. Uma delas é que apenas funciona com entradas binárias, o que significa que entradas de outros tipos devem ser binarizadas na etapa de pré-processamento dos dados. Existem variados esquemas de binarização, que dependem do tipo do dado, e.g., quantidade, tempo, categoria, e de de sua importância para o problema. Esquemas comuns de binarização procuram compatibilizar a distância natural do tipo de dado com a *distância*

de Hamming, de forma que as distâncias sejam proporcionais (ver exemplo na Seção 3.5.2). O esquema de binarização é então um fator crítico no desempenho de classificação (precisão) da rede, pois uma binarização que produza pouca diferença entre entradas distintas poderá causar ambiguidade. Além disto, à medida que o número de padrões diferentes apresentados para treinamento da rede aumenta, mais posições de memória são escritas com "1". Se os dados de treinamento forem ruidosos, a maior parte das posições de memória terá valor "1", fazendo com que os neurônios-RAM produzam "1" como saída e os discriminadores-RAM forneçam predominantemente resposta máxima, causando ambiguidade entre classes e comprometendo a capacidade de classificação da rede. Este efeito é chamado de *saturação* dos neurônios-RAM, e é ilustrado na Figura 2.8. A saturação é o resultado de excesso de treinamento (*overtraining*) na WiSARD.

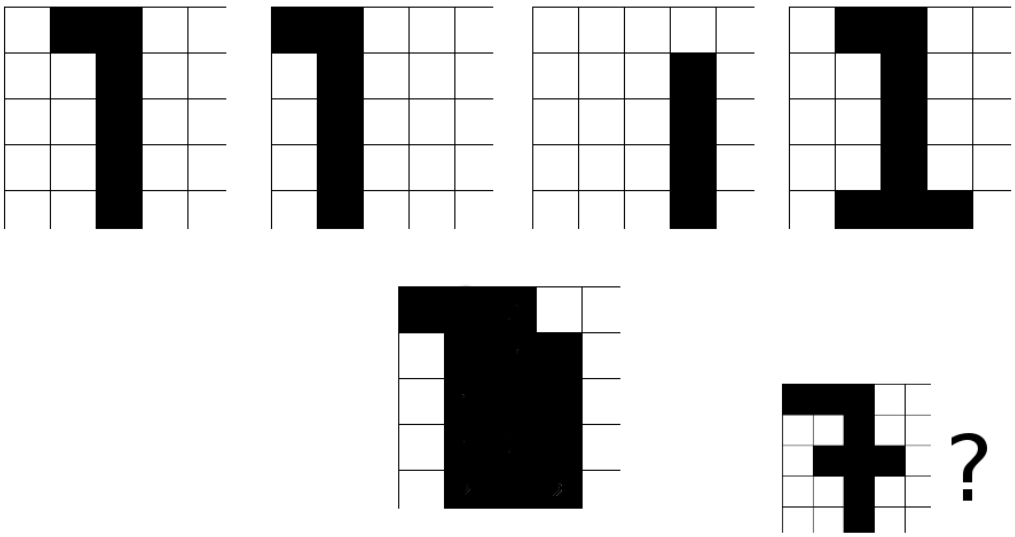


Figura 2.8: Exemplo de saturação em um neurônios-RAM apresentados a alguns padrões para o algarismo "1". O algarismo "7" apresentado para classificação provocará o mesmo grau (máximo) de resposta dos neurônios a quaisquer dos exemplos apresentados, tornando a rede ambígua.

2.4.3 DRASiW e as imagens mentais

O fato de utilizar um mapeamento biunívoco das entradas em memórias para representar a função "aprendida" pela WiSARD, permite que a representação desta função seja obtida através da reversão deste mapeamento. No trabalho de Soares et al. [71], é observado que tal procedimento permite obter exemplares, ou protótipos das classes aprendidas pela WiSARD, e é apresentada a DRASiW, uma extensão da WiSARD voltada à obtenção de tais exemplares. Na DRASiW, as posições de memória passam a armazenar valores inteiros em vez de bits, onde são registradas as quantidades (frequências) de acessos a cada posição. Isto torna possível a construção

das chamadas “imagens mentais”: representações gráficas do conhecimento adquirido pela rede, na forma de um mapa em escala de cinza, no mesmo formato da entrada da rede. Padrões de entrada de natureza visual, como no caso do reconhecimento de caracteres manuscritos, permitem a visualização explícita dos exemplos como imagens aproximadas dos padrões apresentados no treinamento, como ilustrado na Figura 2.9. Os pontos da imagem são mais escuros quanto maior a quantidade de acessos à respectiva posição de memória na rede.

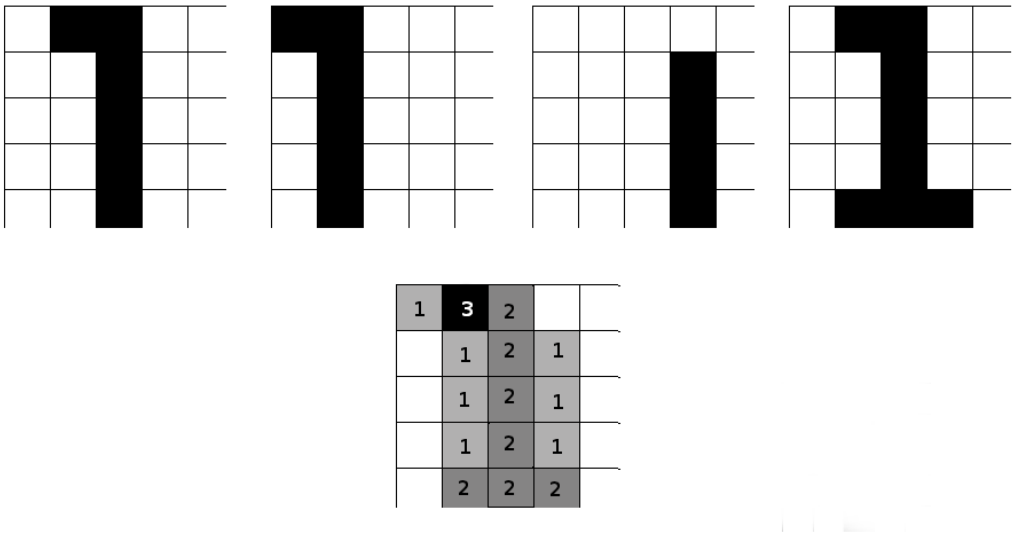


Figura 2.9: Exemplo de imagem mental, com as frequências de acesso de cada entrada. A parte superior mostra exemplos de grade de entrada para imagens representando o caractere "1". A imagem mental apresentada na parte inferior mostra as quantidades de acessos para cada ponto da imagem, conforme registrado na rede. Os pontos com pelo menos um acesso são considerados parte do padrão, levando à saturação da rede. A observação da imagem mental permite identificar sub-padrões mais frequentes (as partes mais escuras), e possivelmente mais relevantes, nos dados apresentados para a rede.

A análise do conteúdo das imagens mentais possibilita uma melhor compreensão do problema a ser resolvido pela rede, levando a melhorias no pré-processamento dos dados e até mesmo à obtenção de regras para classificação, como demonstrado em [72].

2.4.4 *Bleaching e B-bleaching*

A informação da frequência de acessos das posições de memória traz também um outro benefício: facilitar a eliminação de ruído aprendido pela rede. Uma solução para o problema de saturação dos neurônios-RAM consiste na observação de que os padrões representativos de uma classe devem ocorrer mais frequentemente que outros nos exemplos de treinamento. Portanto, a frequência de endereçamento das

posições de memória deve revelar quais partes do padrão armazenado (i.e., subpadrões) são relevantes para o cálculo da similaridade com relação ao conjunto de dados de treinamento, restando apenas encontrar um modo de isolar os subpadrões relevantes dos demais. Este papel é preenchido por uma técnica chamada *bleaching*, apresentada em [73] e explorada em [74]. O objetivo do *bleaching* é eliminar os empates entre discriminadores-RAM, ou seja, controlar a ambiguidade, melhorando a precisão da rede. Isto é feito usando a informação de frequência armazenada pela DRASiW, em conjunto com um filtro seletivo das respostas dos neurônios-RAM. Para esse fim, os valores armazenados nas posições de memória passam a ser considerados como respostas iniciais (r_{ini}) dos neurônios-RAM, não mais limitados a "0" e "1". Define-se uma variável de limiar b, que determina a frequência mínima a ser considerada para a resposta final (r) dos neurônios, obtida da seguinte função:

$$r = \begin{cases} 1, & \text{se } r_{ini} \geq b \\ 0, & \text{do contrário} \end{cases}$$

Em seguida, as respostas r são somadas para obter as respostas dos discriminadores. Começando por $b = 0$, o limiar é incrementado enquanto houver empate nas maiores respostas dos discriminadores. Terminados os empates, é escolhida a classe correspondente ao discriminador com a maior resposta. O processo de *bleaching* é ilustrado na Figura 2.10.

O *bleaching* age diretamente sobre a saturação dos neurônios-RAM, fazendo-os ignorar os subpadrões considerados atípicos, i.e., aqueles que foram apresentados a rede menos que b vezes. Tal procedimento deixa apenas os subpadrões relevantes, resolvendo assim o problema da saturação. Entretanto deve ser notado que se b for muito alto, apenas os subpadrões mais frequentes serão mantidos e a rede perderá capacidade de generalização para variaçoes menores do padrão desejado. Se b for muito baixo, a saturação pode persistir, e junto com esta os empates.

Encontrar o valor ótimo de b requer um procedimento de busca, que pode ser feito por uma variedade de algoritmos, dentre os quais vale destacar:

- *Busca sequencial*: incrementa b em uma unidade até que os empates sejam eliminados com uma confiança c maior que um limiar d;
- *Busca por confiança* [75]: similar à busca sequencial, mas com um incremento variável de b. Ela para ao encontrar o primeiro máximo local de c;
- *Busca binária* [74]: realiza uma busca binária em b, $b \in [1, b_max]$, onde b_max é o maior valor em qualquer posição de memória de qualquer neurônio-RAM do discriminador. É usada a média geométrica no lugar da média aritmética. A

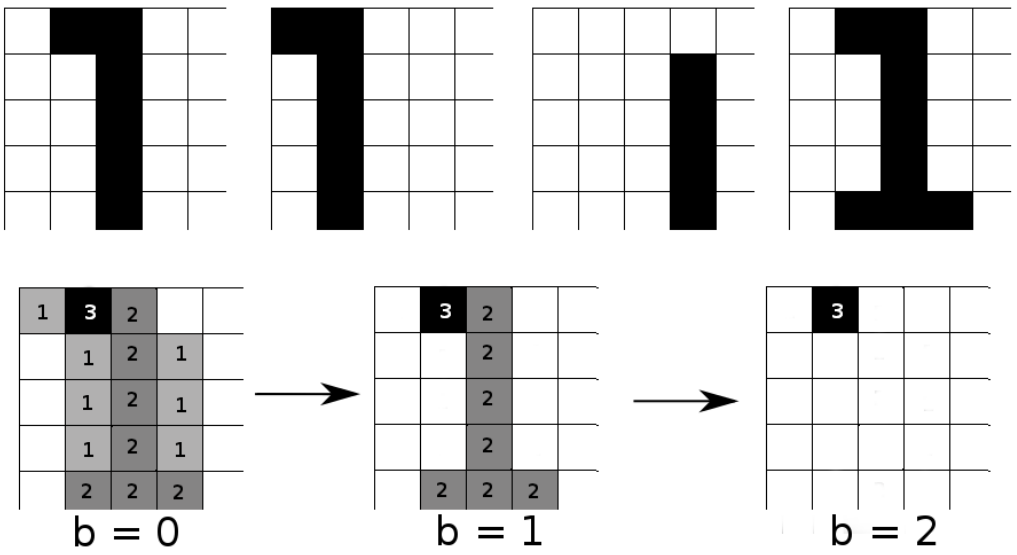


Figura 2.10: Exemplo de *bleaching*, com seu efeito na imagem mental do discriminador. Com $b = 0$, não há *bleaching* e ocorre saturação parcial no discriminador. Com $b = 1$, o bleaching eliminou a saturação, deixando os subpadrões mais frequentes. Com $b = 2$, restou apenas um fragmento do que foi aprendido pela rede e esta perdeu informação relevante.

> busca termina quando é encontrado um valor de b para o qual não há empates e o valor da maior resposta é o mesmo de quando $b = 1$.

Conforme apresentado em [74], a busca binária parece ser a mais rápida dentre as alternativas mencionadas, com mínimo impacto na precisão da rede em relação à busca por confiança, que é a mais precisa. O *bleaching* com busca binária é chamado de *B-bleaching*.

A técnica de *bleaching* pode ser vista como análoga ao processo de poda (*pruning*) aplicado às ANNs tradicionais [76] e a outros modelos de aprendizado de máquina, como as árvores de decisão [77]. No caso das ANNs tradicionais, são removidas sinapses ou neurônios que não impactam ou impactam negativamente na precisão da rede. O objetivo de ambas é minimizar o efeito do ruído aprendido quando não se pode controlá-lo no conjunto de dados de treinamento, sem que a rede sofra perda de sua capacidade de generalização.

Capítulo 3

AS^{2}ABER: Um Anotador de Segmentos Semânticos com Aprendizado Baseado Em Redes Neurais sem Peso

Conforme visto no capítulo 2, a identificação correta das funções conceituais de uma sentença, representadas na forma de segmentos semânticos, e também de suas relações, envolve uma quantidade expressiva de etapas, desde a decomposição do texto até a representação do conhecimento. Cada uma destas etapas pode ser expressa como o conjunto de técnicas e artefatos necessários para sua execução. Cada etapa constitui uma atividade complexa, portanto a sinergia entre as técnicas utilizadas é de vital importância para a qualidade do resultado desejado. Visando obter a melhor combinação de técnicas para alcançar os objetivos deste trabalho, foi implementado o sistema AS^{2}ABER (**A**notador de **S**egmentos **S**emânticos com **A**prendizado **B**aseado **E**m **R**AM), que será descrito neste capítulo. Este sistema tem por objetivo a identificação e anotação de segmentos semânticos e de suas relações em texto livre, escrito em linguagem natural, tendo como princípio de funcionamento o aprendizado de padrões morfossintáticos através de um conjunto de técnicas, em especial as *Redes Neurais sem Peso*, também conhecidas como *Redes Neurais Baseadas em RAM* (Seção 2.4.1).

3.1 Estrutura geral

3.1.1 Características do sistema

O funcionamento do sistema é dividido em duas fases distintas: Treinamento e Extração. Suas entradas e saídas dependem da fase a ser executada. Durante a fase

de treinamento, são apresentados ao sistema as sentenças em linguagem natural diretamente extraídas do texto, junto a uma versão das mesmas com anotações manuais dos segmentos semânticos. Os padrões a serem aprendidos são obtidos a partir destas duas entradas. Ao fim desta fase, o sistema produz as seguintes saídas:

- Um modelo neural WiSARD treinado com os padrões aprendidos para cada classe de segmento.
- Um conjunto de pares hierárquicos (pai, filho) de classes de segmento para os exemplos analisados.
- Um conjunto de tabelas (classe segmento -> classe sintática) para os alinhamentos ocorridos nos exemplos analisados.
- Um conjunto de padrões morfológicos (seção 3.5.1) para cada classe de segmento.
- Uma lista de posições relativas ocupadas por cada segmento em relação à sentença em análise.
- Um conjunto de tamanhos mínimos e máximos (em palavras) para cada classe de segmento.

Para a fase de extração, são apresentadas ao sistema as sentenças em linguagem natural das quais se desejam obter os segmentos, junto a todas as saídas da fase de treinamento. Além destes, o sistema também recebe um conjunto de regras de relacionamento entre classes de segmentos, construído manualmente. Ao fim desta fase, o sistema produz para cada sentença uma lista com os segmentos extraídos e suas respectivas classes, e um grafo de relacionamento entre os segmentos.
Tanto na fase de treinamento quanto na de extração, as reivindicações de patente são apresentadas ao sistema em sua forma pura: sentenças em linguagem natural obtidas da seção de reivindicações presente em qualquer documento de patente.

O sistema usa as seguintes ferramentas externas:

- **mWANN-Tagger** [31]: para análise morfológica (POS-tagging) das sentenças;
- **LX-Parser** [78]: para análise sintática (parsing) das sentenças;
- **NLTK** [79]: para operações em árvores sintáticas e de segmentos semânticos.

3.1.2 Arquitetura

A arquitetura adotada no sistema foi baseada no modelo de processamento em linha de montagem (*pipeline*), onde a saída de uma etapa serve de entrada para as etapas posteriores. As Figuras 3.1 e 3.2 mostram uma visão geral do fluxo de operações do sistema, dividido em suas duas fases de operação. As Figuras 3.3, 3.4, 3.5 e 3.6 detalham os módulos de alinhamento sintático-semântico, treinamento do classificador, extração de padrões morfológicos e extração e classificação de segmentos respectivamente.

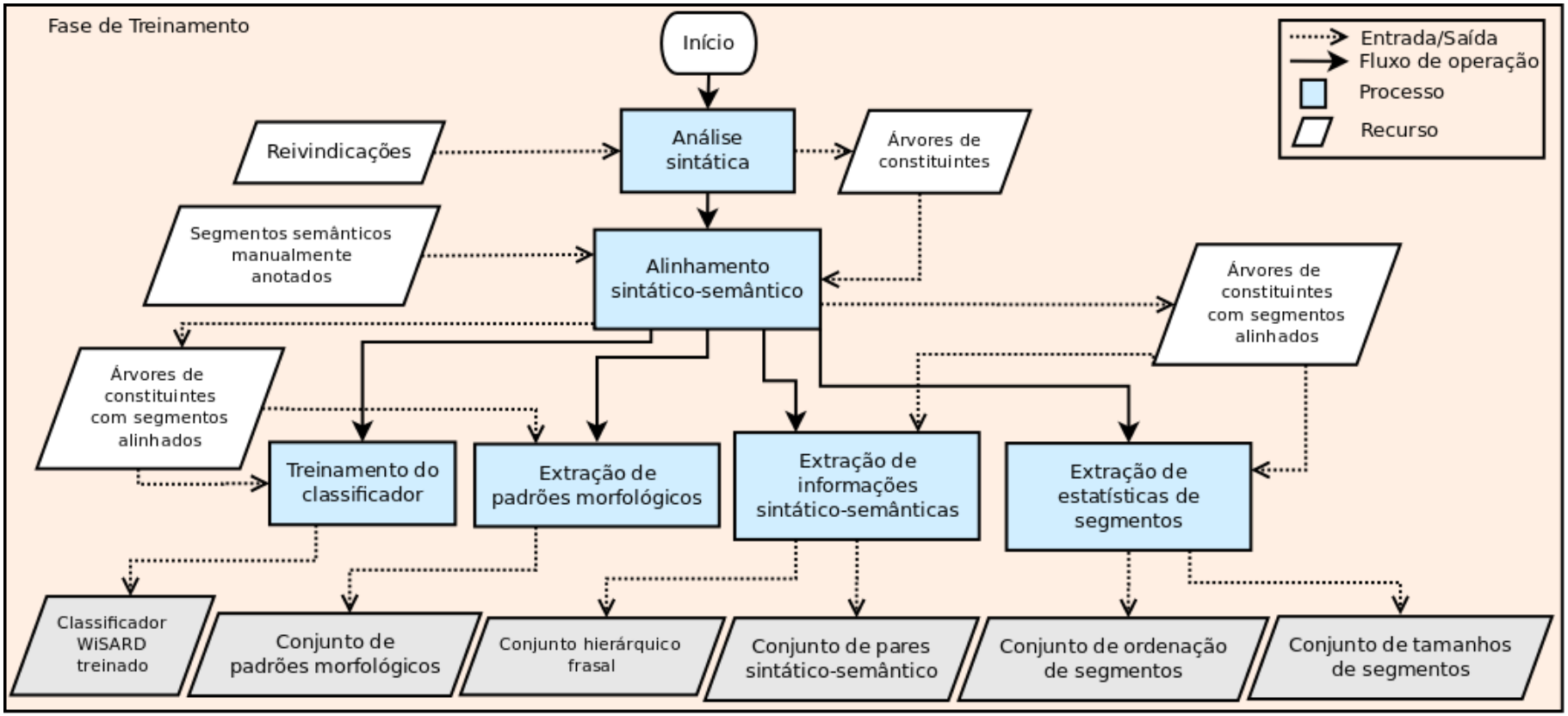


Figura 3.1: Fluxo de operações do sistema para a fase de treinamento. Cada operação realiza a leitura de um conjunto de entradas e produz um conjunto de saídas. Todas as saídas finais da fase de treinamento são utilizadas na fase posterior: extração e classificação de segmentos.

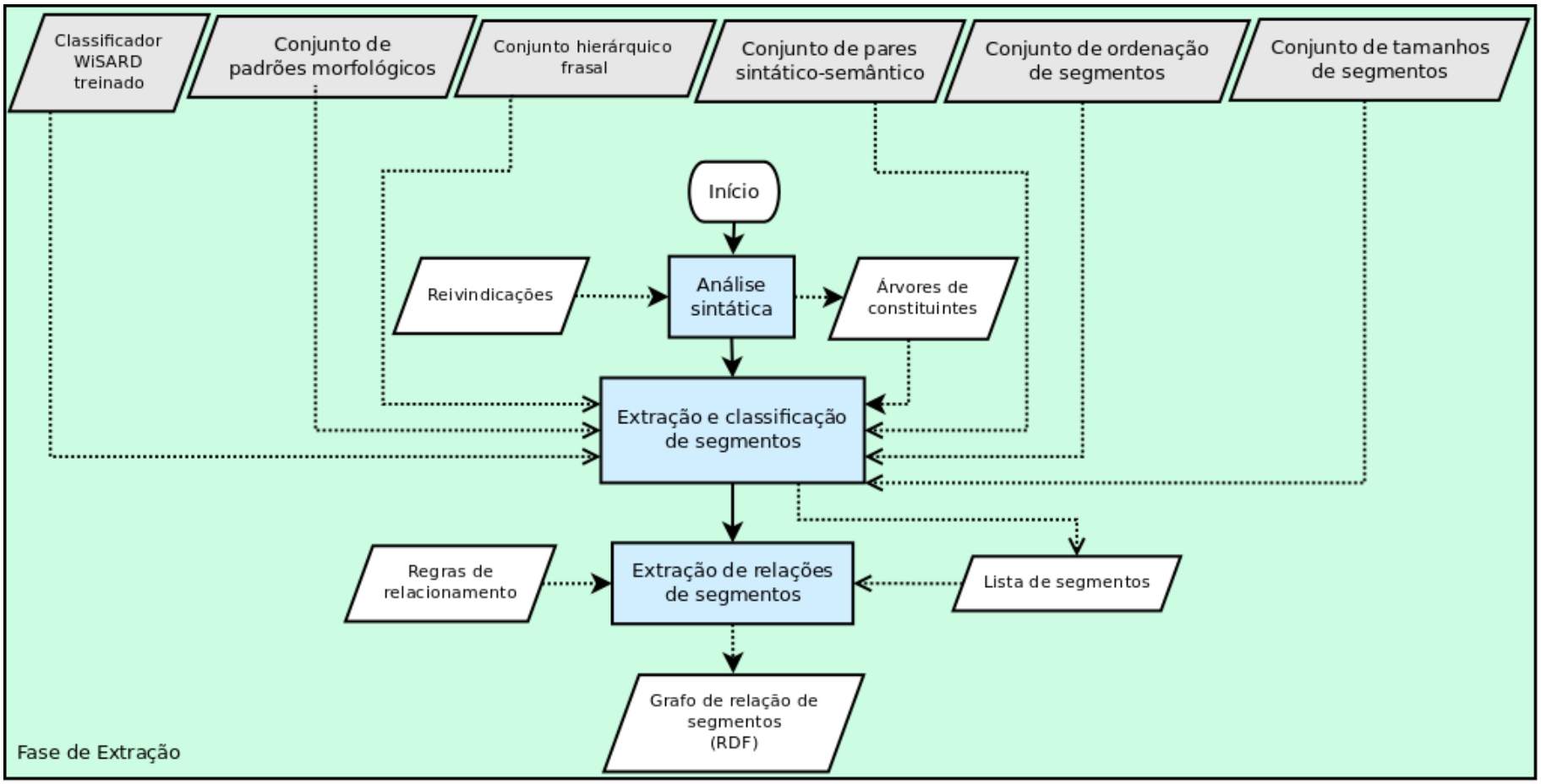


Figura 3.2: Fluxo de operações do sistema para a fase de extração. Cada operação realiza a leitura de um conjunto de entradas e produz um conjunto de saídas. Todas as saídas finais da fase de treinamento são exibidas no topo.

3.2 Obtenção dos Documentos de Patente

Os documentos de patente utilizados neste trabalho foram obtidos através do serviço de consulta pública a patentes do INPI[1], o escritório de patentes brasileiro. Tal serviço possibilita a consulta dos documentos de pedidos e de patentes já concedidas, através de diversos atributos, como título, data do pedido, palavras no resumo, entre outros. O serviço é gratuito, mas limitado na quantidade de documentos que podem ser obtidos em um certo intervalo de tempo. Além disto, nem todos os documentos estão disponíveis integralmente. A maior parte dos documentos registrados antes de 2006 contém apenas o resumo da patente.

Todos os documentos de patente disponibilizados pelo INPI através do serviço de consulta pública estão na forma de arquivos PDF, resultantes da digitalização (*scanning*) dos formulários gerados no processo de concessão (e.g., o formulário de pedido de patente). Esta digitalização é feita sem a aplicação de OCR[2] (Reconhecimento Óptico de Caracteres), resultando em imagens sem nenhum texto associado. Isto significa que, para obter o texto dos documentos, é necessário efetuar o OCR em primeiro lugar. Entretanto, a utilização de OCR em documentos já digitalizados apresenta algumas dificuldades:

- Falhas na digitalização original (e.g., posicionamento, contraste) não podem ser corrigidos;
- Ruídos gerados pela compressão aplicada na digitalização (e.g., JPEG) dificultam a diferenciação de alguns caracteres, como "1" (numeral UM) e "l" (letra L minúscula) e aqueles os que possuem acentos;
- Ferramentas livres para OCR (i.e., não atreladas a um *scanner*) possuem menos recursos de ajuste automático dos algoritmos em relação às embutidas nos *scanners*.

Considerando estes fatores, foi utilizada a ferramenta livre Tesseract-OCR [80] para efetuar OCR nos documentos obtidos. O Tesseract-OCR foi escolhido por ser a referência de qualidade entre as ferramentas de OCR livres e por permitir ajustes manuais conforme o documento a ser analisado. Após o OCR, foi feita a correção manual dos textos extraídos para eliminar a maioria dos erros, preservando a estrutura original dos textos. Este é um processo lento e laborioso, sendo inviável sua aplicação para grandes quantidades de documentos. Entretanto, espera-se que este problema seja minorado com a futura transição para um sistema de depósito digital de patentes[3], planejado pelo INPI.

[1] Instituto Nacional de Propriedade Industrial
[2] *Optical Caracter Recognition*
[3] http://epatentes.inpi.gov.br/modulo2/edeposito/

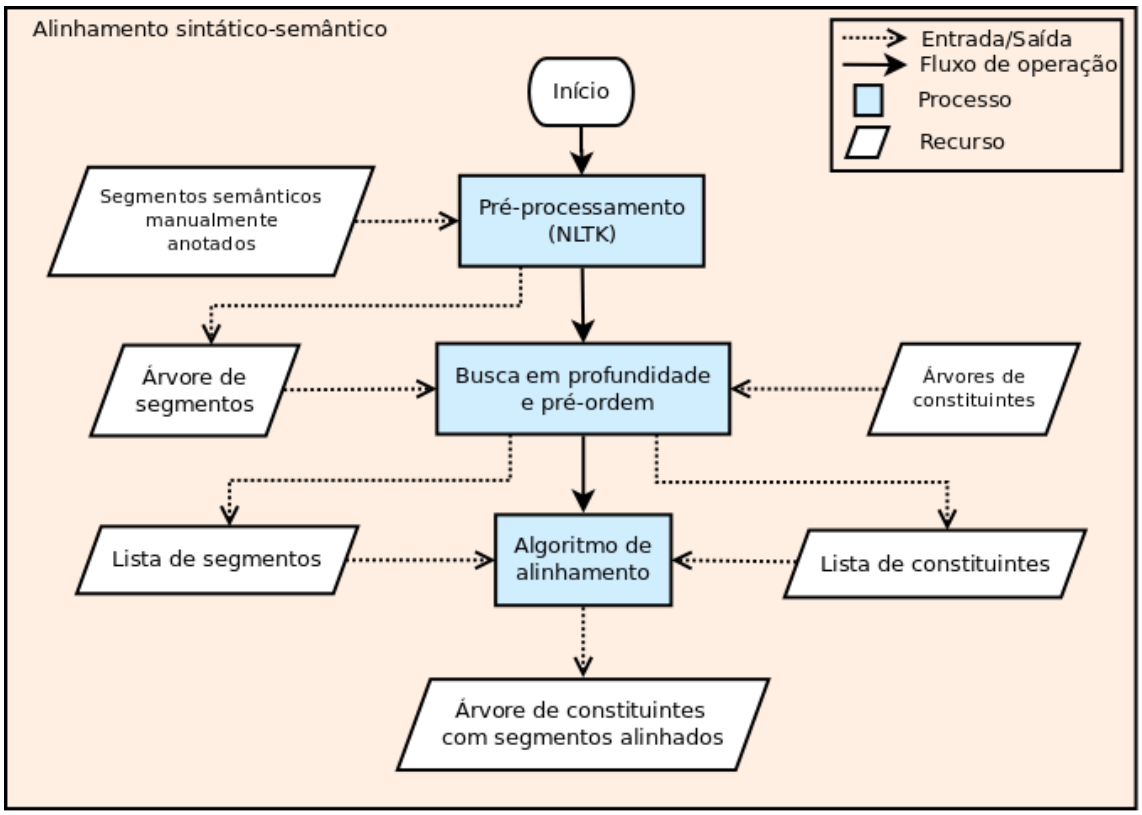


Figura 3.3: Fluxograma do módulo de alinhamento sintático-semântico.

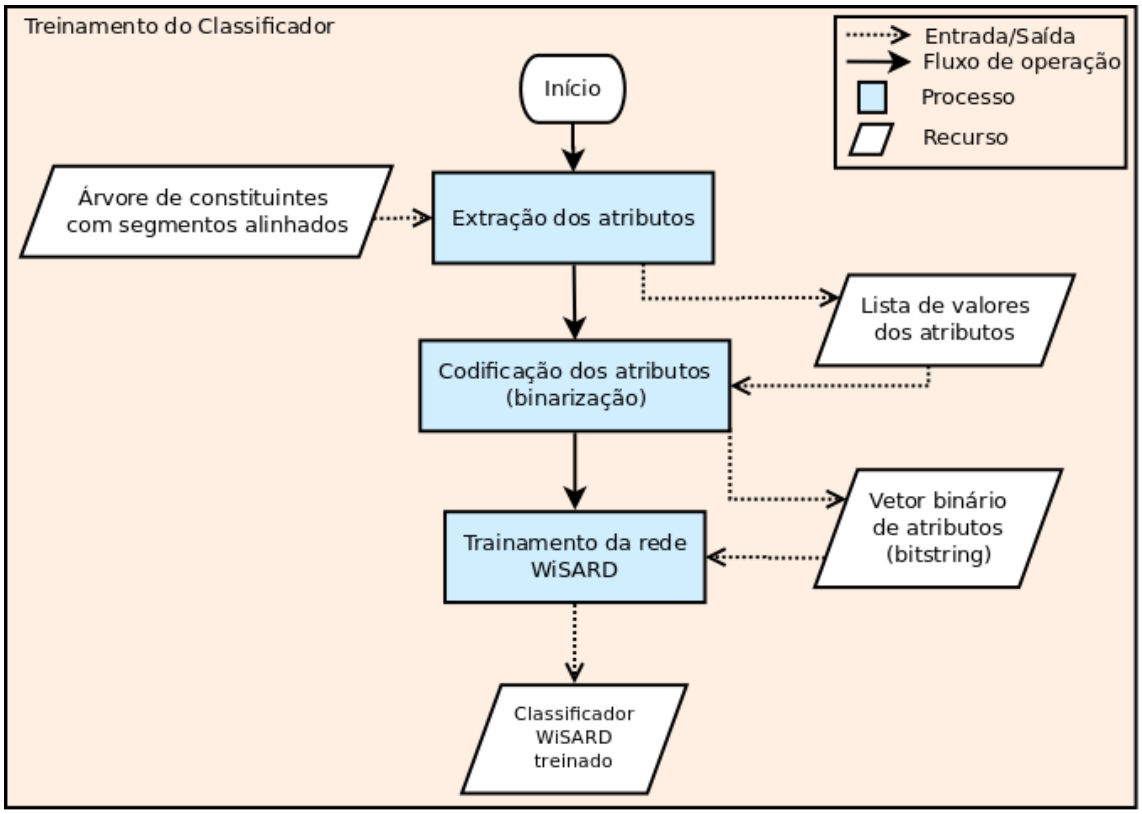


Figura 3.4: Fluxograma do módulo de treinamento do classificador.

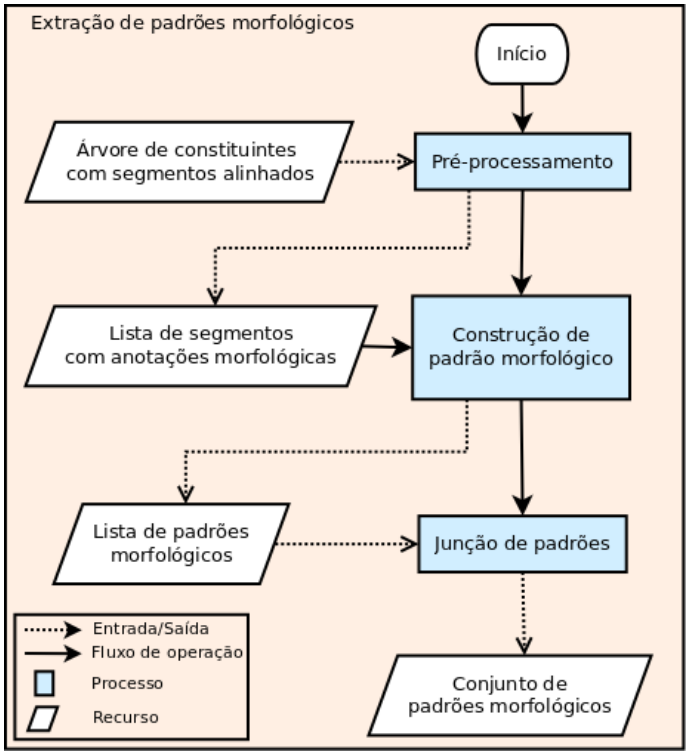


Figura 3.5: Fluxograma do módulo de extração de padrões morfológicos.

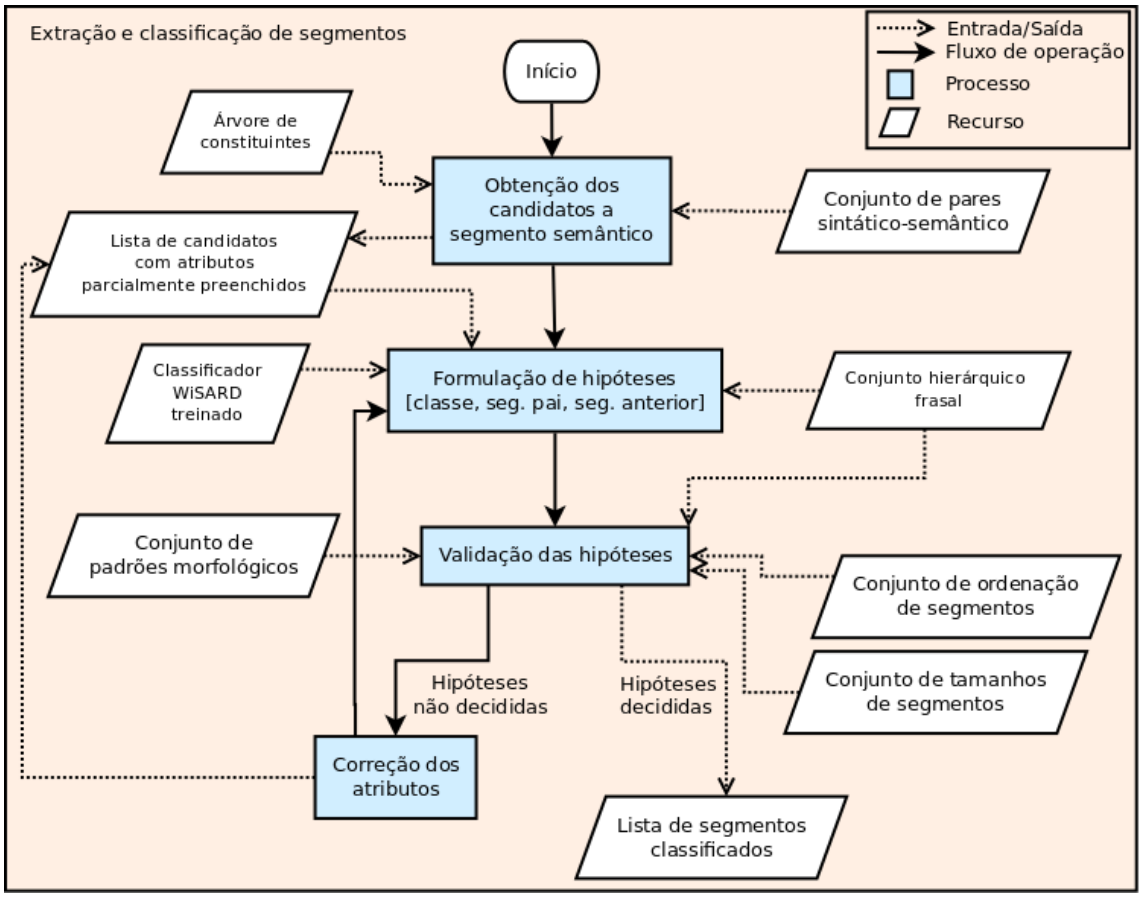


Figura 3.6: Fluxograma do módulo de extração e classificação de segmentos.

3.3 Extração e análise do texto de reivindicações

Um vez obtido o texto do documento de patente, o próximo passo é isolar a seção de reivindicações que será processada pelo sistema. Esta tarefa é conhecida como *Spotting*. Neste trabalho o *spotting* é feito manualmente, pois pertence ao escopo de processamento de documentos, que não é coberto neste trabalho. Para alguns documentos é possível obter apenas a seção de reivindicações através do serviço de consulta, eliminando a necessidade do *spotting*.

Com as reivindicações isoladas, é possível fazer o treinamento e a extração dos segmentos semânticos. Para ambas as fases, o primeiro passo é realizar a tokenização e análise morfológica (*POS-tagging*), seguida da análise sintática 2.3.1 (*parsing*) de cada reivindicação. A fim de limitar o escopo do trabalho a um conjunto típico de reivindicações, considera-se que há apenas uma reivindicação por sentença.

O objetivo da análise sintática é obter a árvore de constituintes (Figura 2.3), de cada sentença, que será usada nas etapas posteriores para indicar (no treinamento) ou ajudar na identificação de quais constituintes representam os segmentos semânticos contidos na sentença. Para esta tarefa foi utilizada a ferramenta *LX-Parser* [78], um analisador baseado em PCFG [35]. O LX-Parser foi escolhido em função da sua pronta disponibilidade no site do desenvolvedor, e de produzir as árvores no formato desejado (*Penn Treebank* [81]).

Entretanto, a análise sintática para textos em português apresenta algumas dificuldades, principalmente em relação à qualidade das árvores resultantes. Infelizmente o estado da arte de analisadores sintáticos para o português ainda deixa a desejar, quando comparado com aqueles feitos para a língua inglesa, geralmente usados como referência para esta tarefa. Isto ocorre devido a alguns fatores como a menor quantidade de material anotado disponível para esta língua, em relação ao inglês por exemplo, que conta com uma fartura de recursos linguísticos. Uma alternativa ao analisador sintático foi inicialmente proposta como solução, escolhendo-se um *analisador sintático raso* (*chunker*) para prencher este papel. Entretanto, esta alternativa foi descartada pois o *chunker* produz árvores com no máximo dois níveis de profundidade, não capturando o detalhe necessário para identificar segmentos pequenos. Decomposições da árvore gerada pelo *chunker* também foram experimentadas, através da elaboração de um *chunker* recursivo. Esta última alternativa produziu resultados melhores para um pequeno conjunto de sentenças, mas piores em maioria.

Após uma série de experimentos e tentativas de reduzir os erros de parsing e melhorar a qualidade das árvores resultantes, foi constatado que uma grande parte dos erros era causada por erros na etapa anterior, a análise morfológica. A solução adotada foi substituir o analisador morfológico embutido no LX-Parser, chamado

LX-Tagger, pelo *mWANN-Tagger* [31], que possui acurácia consideravelmente mais alta. Como os LX-Tagger e o mWANN-Tagger utilizam padrões diferentes de anotação, foi necessária a elaboração de um mapeador da saída do segundo para o primeiro, de forma que sua saída ficasse compatível com a entrada esperada pelo LX-Parser. Uma tabela com os mapeamentos pode ser encontrada no Apêndice A.1. Com a integração do mWANN-Tagger, as árvores obtidas do LX-Parser passaram a ter a mínima qualidade necessária para a operação do sistema.

Um outro fator que dificulta a análise sintática é o tamanho da sentença. Sentenças muito longas, com mais de 40 palavras, causam considerável redução de acurácia em *parsers* PCFG. Entretanto, reivindicações costumam ser longas, algumas vezes com mais de 200 palavras. Uma estratégia para contornar este problema é apresentada no trabalho de Yang e Soo [3], consistindo em quebrar a sentença em partes independentes sintaticamente. A abordagem escolhida para este trabalho foi a de utilizar a sentença inteira e deixar que os padrões aprendidos pelo sistema guiem a segmentação. Esta escolha foi feita em função da dificuldade em elaborar heurísticas abrangentes para isolar termos de domínio e evitar que eles sejam divididos em partes diferentes da sentença, considerando que eles são um alvo do modelo de segmentação utilizado (Seção 3.4). A Figura 3.7 ilustra o resultado esperado da análise sintática na reivindicação "*Dispositivo de acordo com a reivindicação 1, caracterizado pela caixa blindada 1 ser produzida em material rígido e resistente a impacto*".

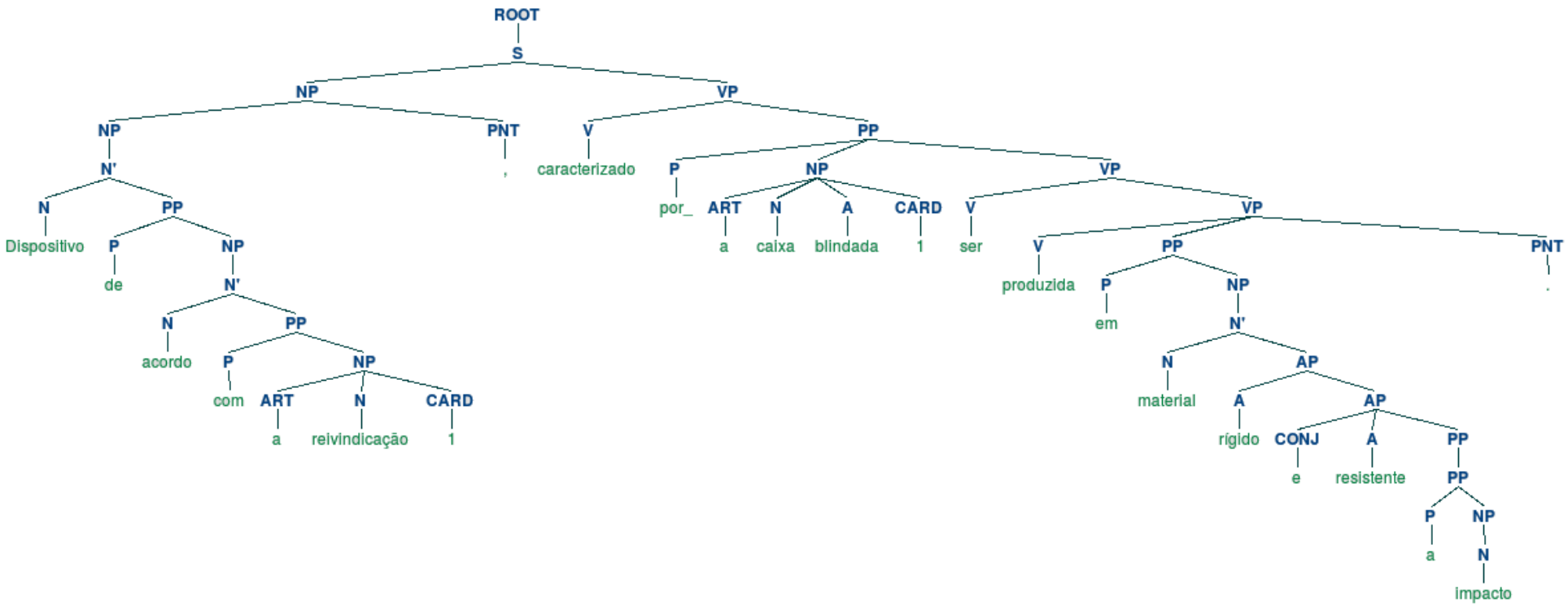


Figura 3.7: Árvore de constituintes para a reivindicação "*Dispositivo de acordo com a reivindicação 1, caracterizado pela caixa blindada 1 ser produzida em material rígido e resistente a impacto*". Os nós não terminais são marcados com suas respectívas classes sintáticas, como "NP": Noun Phrase (sintagma nominal) e "VP": Verb Phrase (sintagma verbal). As folhas são marcadas com as classes gramaticais, como "N": Noun (substantivo) e "A": Adjetivo.

3.4 Modelo de segmentação semântica

O sistema usa como modelo de segmentação semântica uma pequena ontologia de segmentos, obtida a partir das anotações fornecidas na fase de treinamento e complementada pelas regras de relacionamento fornecidas na fase de extração. Os nomes das classes e seus relacionamentos variam conforme a aplicação, mas os atributos são fixos e iguais para todas as classes. O sistema de anotação utilizado é baseado no formato Penn Treebank [81], sendo portanto de natureza hierárquica. Desta forma, o único tipo de relacionamento disponível é o parte-todo. A sentença abaixo e sua segmentação exemplificam o formato de marcação utilizado:

Blindagem protetora de acordo com a reivindicação 1, caracterizada pelo fato de ditos painéis estarem solidamente unidos entre si ao longo de suas bordas.

```
(ROOT
    (ASSUNTO_PAT Blindagem protetora)
    (REF de acordo com
        (REF_REIVIND a reivindicação
            (REF_REIVIND_NUM 1)
        )
    ) ,
    (CARACT_ASSUNTO caracterizada por_ o fato de
        (OBJ_PAT ditos painéis
            (CARACT_OBJ estarem solidamente unidos entre si a_ o longo de
                (OBJ_PAT suas bordas)
            )
        )
    )
)
```

A ontologia usada para segmentação das reivindicações de patentes é descrita à seguir.

3.4.1 Ontologia de segmentos

Classes

- *ASSUNTO_PAT*: Assunto da patente e principal objeto de proteção legal da reivindicação. Possivelmente a informação mais importante em uma reivindicação, pois define o tópico a ser detalhado no restante da sentença. Ex: "[Pistão hidráulico 1] acionado através do fluido hidráulico e ...".
- *REF*: Referência a uma reivindicação no mesmo documento ou a outra patente. Indica que o tópico da reivindicação está relacionado a algo não descrito na

mesma. Ex: “Dispositivo [de acordo com a reivindicação 1] ...”.

- *REF_REIVIND*: Referência explícita a uma reivindicação do mesmo documento. Indica que a reivindicação sendo analisada é dependente daquela que foi referenciada. Ex: “Dispositivo de acordo com a [reivindicação 1]”.

- *REF_REIVIND_NUM*: Número da reivindicação referenciada. Ex: “Dispositivo de acordo com a reivindicação [1]”.

- *CARACT_ASSUNTO*: Caracterização do tópico da reivindicação. Detalha o objeto de proteção legal, diferenciando-o de outros similares. Ex: “Este dispositivo [é caracterizado por se adaptar nos frascos ...]”.

- *OBJ_PAT*: Declaração de objeto secundário, que caracteriza o tópico da reivindicação. Ex: “...compreendendo [um conjunto de painéis de blindagem] ...”.

- *CARACT_OBJ*: Caracterização de um objeto secundário. Ex: “...uma superfície parabólica refletiva 1, [que utiliza a base de fixação 2]”

- *ILUST_REF*: Referência a uma ilustração no documento, geralmente especificando um objeto ou o tópico da reivindicação. Ex: “Pistão hidráulico [1] acionado através do fluido hidráulico e ...”.

Atributos

- *Frequência de classes gramaticais*: Vetor onde cada posição representa uma classe gramatical (*POS tag*) diferente, totalizando 10 posições. Cada posição é preenchida com o número de vezes que a classe correspondente aparece no segmento, contando todas as folhas.
  Ex: $\frac{O}{ART}\ \frac{cachorro}{NN}\ \frac{gordo}{ADJ}\ \frac{estava}{V}\ \frac{feliz}{ADJ}\ \frac{porque}{CONJ}\ \frac{estava}{V}\ \frac{comendo}{V}$. Vetor: (NN, ART, ADJ, V, PREP, CONJ) -> [1, 1, 2, 3, 0, 1].

- *Ordem de classes gramaticais*: Vetor igual ao anterior, onde cada posição é preenchida com a ordem de aparição da classe correspondente no segmento, considerando apenas a primeira aparição de cada classe.
  Ex: Para a frase anterior (NN, ART, ADJ, V, PREP, CONJ) -> [2, 1, 3, 4, 0, 5].

- *Número de palavras*: Número de palavras do segmento.

- *Formato título*: Se o segmento possui formato de título, ou seja, tem todas as palavras em maiúsculas ou capitalizadas.

- *Classe sintática*: *Tag* sintática do nó da arvore sintática da sentença, após alinhamento (seção 3.5.1), que melhor representa o segmento.
- *Classe sintática pai*: *Tag* sintática do nó pai daquele onde o segmento foi alinhado.
- *Classe semântica pai*: Se o segmento está contido em outro, esta é a classe deste último, do contrário é vazia (nenhuma classe).
- *Classe semântica anterior*: Se o segmento não é o primeiro da sentença, esta é a classe do vizinho anterior, se for o primeiro, é vazia (nenhuma classe).

Estes atributos foram selecionados após observação cuidadosa de diversos textos e intuições sobre a correlação entre estruturas linguísticas e seus significados, tais como apresentados por Chomsky [34] e Jackendoff [42]. Os atributos deveriam permitir um mapeamento próximo daquele feito por um ser humano, assumindo apenas a presença da informação sintática, servindo portanto a qualquer domínio de conhecimento.

Experimentos feitos em trabalhos anteriores [82, 83] apontaram para uma considerável relevância da informação de constituintes sintáticos, em especial a ordem relativa de constituintes e classes gramaticais na sentença e o tipo e quantidade de classes gramaticais presentes. Os demais atributos foram selecionados a partir da observação de recursos linguísticos utilizados em textos formais e declarativos, como é o caso dos documentos de patente.

Regras de relacionamento

As regras de relacionamento (Tabela A.2, Apêndice A) determinam como as diferentes classes de segmento podem estar ligadas entre si além das relações já contidas nas anotações, visto que estas últimas são apenas de natureza hierárquica. Elas definem relações baseadas em sequências de segmentos, levando em conta a ordem do discurso. O resultado da aplicação das regras é um conjunto de triplas (sujeito, predicado, objeto), que pode ser representado na forma de um *grafo de relacionamento semântico*, onde os sujeitos e objetos são nós e os predicados são arestas. Tal grafo serve a mais de um propósito: (*i*) auxiliar a visão da estrutura de um documento, através da concatenação dos grafos de suas reivindicações; (*ii*) facilitar a comparação entre diferentes documentos através de algoritmos de alinhamento (*matching*) de grafos de conhecimento; (*iii*) facilitar a criação ou complementação de uma ontologia de reivindicações de patentes.

Exemplo: Sequência: (ASSUNTO_PAT → REF → REF_REIVIND) **então** Tripla: (sujeito[ASSUNTO_PAT], predicado[de acordo com], objeto[REF_REIVIND])

3.5 Treinamento do modelo

Ao longo da fase de treinamento, cada reivindicação passa por uma série de etapas, onde cada uma é responsável por obter uma informação diferente a respeito dos segmentos. Estas etapas são descritas nas seções 3.5.1 a 3.5.2.

3.5.1 Mapeamento sintático-semântico

Alinhamento de árvores: estrutura frasal X semântica

Conforme visto na seção 3.4, o modelo de segmentação semântica adotado é de natureza hierárquica. Portanto, para cada sentença podem ser obtidas duas representações em árvore distintas: a *árvore de constituintes sintáticos* e a *árvore de segmentos semânticos*. Sabendo haver uma correspondência dos nós da árvore de segmentos para alguns nós da árvore de constituintes (Seções 2.3.1 e 2.3.2), define-se o processo de *alinhamento sintático-semântico* como o conjunto de operações necessárias para obter um mapeamento dos nós da árvore de segmentos em nós da arvore de constituintes, tal que a frase contida no nó constituinte seja a mais próxima possível daquela contida no nó semântico. Um alinhamento perfeito entre nós ocorre quando as frases são iguais. A Figura 3.8 ilustra uma árvore de segmentos semânticos para mesma sentença da Figura 3.7, e a Figura 3.9 ilustra o resultado do alinhamento das duas árvores. O Algoritmo 1 descreve o procedimento para alinhamento das árvores.

Algoritmo 1 Algoritmo de alinhamento sintático-semântico

```
1: lista_de_segmentos ← visita nós da árv. segmentos em profundidade e pré-ordem
2: lista_de_constituintes ← visita nós da árv. constituintes em profundidade e pré-ordem
3: ultima_posicao ← 0
4: para cada nó_segmento em lista_segmentos faça
5:     posicao ← ultima_posicao
6:     para cada nó_constituinte em lista_constituintesa partir de posicao faça
7:         se (similaridade(no_segmento.frase, no_constituinte.frase) > LIMIAR) então
8:             marca_no_alinhado(no_constituinte)
9:             ultima_posicao ← posicao
10:        fim se
11:        posicao ← posicao + 1
12:    fim para
13: fim para
```

A similaridade entre as frases contidas no nó semântico e no nó constituinte e calculada através de sobreposição de strings. O valor resultante pertence ao intervalo $[0, 1]$, sendo o valor 0 indicativo de nenhuma sobreposição e o valor 1 indicativo de sobreposição total.

Devido a falhas na geração da árvore de constituintes, frases que deveriam estar em um nó podem ficar separadas em subárvores adjacentes, fazendo com que não possam ser alinhadas por este método. Valores altos de $LIMIAR$ fazem com que apenas pequenas diferenças entre o conteúdo do segmento e do nó sintático sejam toleradas, como é o caso de artigos ou preposições em falta ou excesso nas bordas da frase. Valores mais baixos permitem maior tolerância aos erros de *parsing*, mas induzem a produção de ruído, i.e., alinhamentos inconsistentes. O valor de $LIMIAR$ usado para este trabalho é constante e igual a 0.9, valor obtido após testes para minimização do ruído.

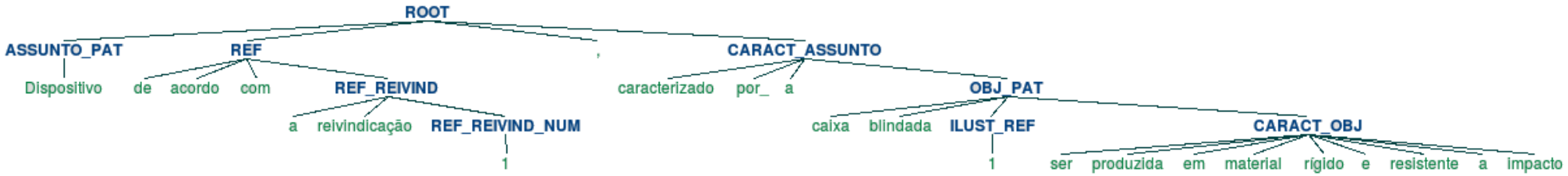


Figura 3.8: Árvore de segmentos semânticos.

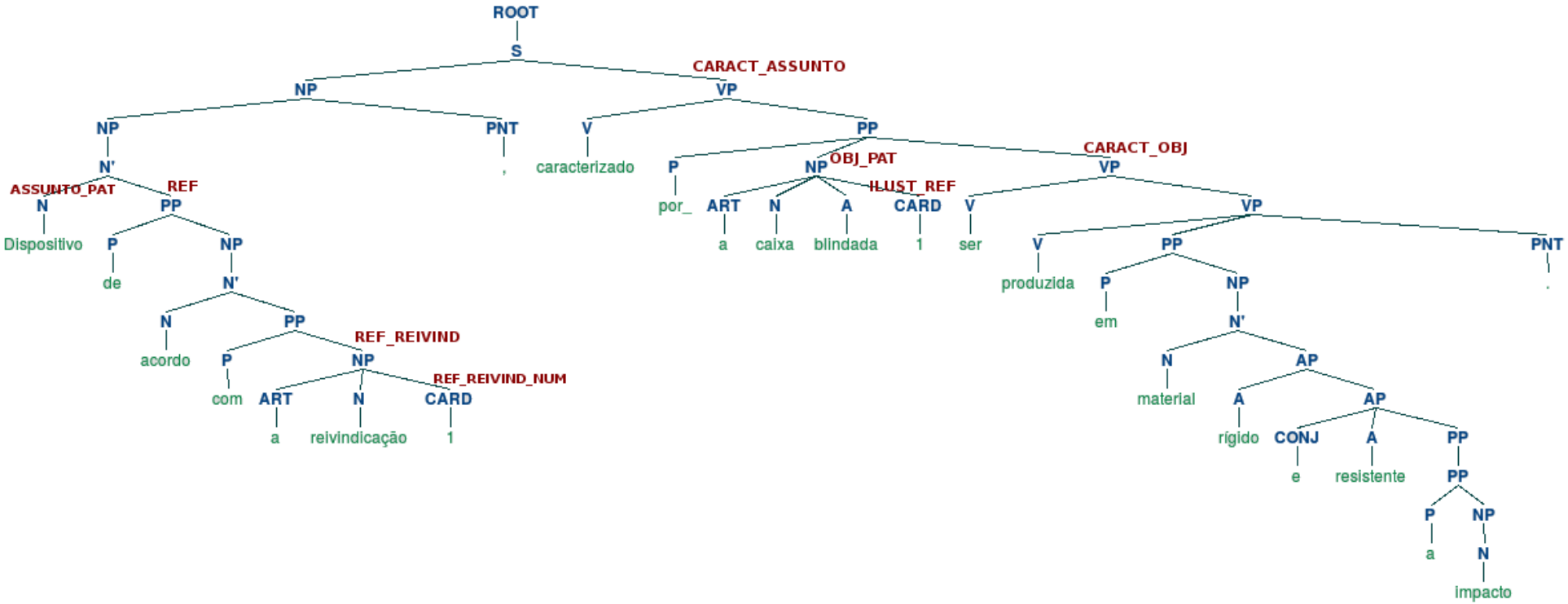


Figura 3.9: Alinhamento sintático-semântico.

Conjunto hierárquico frasal

Durante o alinhamento sintático-semântico, as árvores de segmentos são completamente visitadas e com isso é possível observar todos os pares hierárquicos (parte, todo) que formam o conjunto de relações entre as frases analisadas. Este conjunto de pares pode ser utilizado para validar possibilidades de classificação de segmentos ou simplesmente para limitar o espaço de classes a ser considerado, quando um dos elementos do par é conhecido.

Exemplo: no trecho de reivindicação anotada abaixo

```
(ROOT (ASSUNTO_PAT Pistão hidráulico (ILUST_REF 1))
      (CARACT_ASSUNTO aplicável a
        (OBJ_PAT uma asa
            (ILUST_REF 13)
         de um pulverizador agrícola)
      )
      ...
```

os pares (ASSUNTO_PAT, ILUST_REF), (CARACT_ASSUNTO, OBJ_PAT) e (OBJ_PAT, ILUST_REF) podem ser observados.

Conjunto de pares sintático-semânticos

Como resultado do alinhamento, podem ser observados todos os possíveis pares (classe de segmento, classe sintática) dos exemplos de treinamento. Este conjunto de pares também pode ser utilizado para validar possibilidades de classificação de segmentos ou limitar o espaço de classes a ser considerado. Como a árvore de constituintes é também obtida no início da fase de extração, o segundo elemento do par será sempre conhecido.

Exemplo: no alinhamento apresentado na Figura 3.9, os pares obtidos são: (ASSUNTO_PAT, N'), (REF, PP), (REF_REIVIND, NP), (REF_REIVIND_NUM, CARD), (CARACT_ASSUNTO, VP), (OBJ_PAT, NP), (ILUST_REF, CARD), (CARACT_OBJ, VP).

Conjuntos de ordem e tamanho dos segmentos

Outras duas informações importantes sobre os segmentos dizem respeito à ordem em que os mesmos ocorrem nas sentenças e seus tamanhos em palavras. A ordem é especialmente importante nas reivindicações, pois o discurso utilizado é declarativo e portanto tópicos e objetos precisam ser declarados antes de serem especificados ou referenciados. Algumas classes podem ter tamanhos típicos em um intervalo bastante limitado, como é o caso das referências à ilustrações, que apresentam entre

uma e duas palavras (e.g., "Fig. 01").
Ambos os conjuntos são armazenados na forma de mapas (classe de segmento → [ordenações]) e (classe de segmento → [tamanhos]), onde *ordenações* e *tamanhos* são coletados para cada segmento dos exemplos de treinamento.

Exemplo: no alinhamento apresentado na Figura 3.9, as ordenações e tamanhos obtidos são:
Ordenações: ASSUNTO_PAT → [0], REF → [1], REF_REIVIND → [2], REF_-REIVIND_NUM → [3], CARACT_ASSUNTO → [4],OBJ_PAT → [5], ILUST_-REF → [6], CARACT_OBJ → [7].
Tamanhos: ASSUNTO_PAT → [1], REF → [6], REF_REIVIND → [3], REF_-REIVIND_NUM → [1], CARACT_ASSUNTO → [15],OBJ_PAT → [4], ILUST_-REF → [1], CARACT_OBJ → [9].

Padrões morfológicos

Para algumas classes de segmento, há construções frasais que são bastante frequentes, envolvendo as mesmas palavras, ou palavras com a mesma classe gramatical. Exemplos típicos incluem as frases começadas por "caracterizado por", que explicitam uma caracterização do tópico ou outro objeto da patente, ou a construção "[substantivo] [adjetivo] [preposição]" que frequentemente é usada para declarar o tópico, como exemplo em "Tampa inviolável para ...". Identificar corretamente tais padrões pode auxiliar ou ser usado como fator decisivo na classificação de um segmento.

Com o objetivo de capturar estes padrões, foi elaborado um método para representar e generalizar as instâncias de construções frasais encontradas nos segmentos, inspirado no trabalho de Pantel e Pennacchiotti [9]. Seu funcionamento pode ser resumido em três operações principais:

1. Construção: obtém os radicais das primeiras N palavras do segmento e suas respectivas classes gramaticais.

2. Junção: tenta unir duas instâncias distintas. Palavras diferentes na mesma posição com a mesma classe gramatical, são generalizadas para a notação "*". Segmentos com classes gramaticais diferentes na mesma posição não podem ser unidos.

3. Casamento: verifica se dois segmentos se encaixam em um mesmo padrão. O casamento é bem sucedido em caso positivo e mal sucedido do contrário.

Exemplo: para os segmentos "caracterizado pela forma do contato" e "caracterizado por um suporte rígido" os padrões construídos, após a tokenização e análise

morfológica, são respectivamente: "caracter/V por_/PREP a/ART" e "caracter/V por/PREP um/ART". Após a junção, o padrão resultante é "caracter/V por/PREP */ART". O segmento "caracterizado pelo *(por + o)* tubo curvado" casa com este padrão, mas não o segmento "caracterizado por todos os botões ...".

Um padrão é construído para cada nó visitado nas árvores de segmentos, sendo tentada sua junção a todos os outros padrões obtidos da mesma forma. Uma junção bem sucedida produz um padrão mais genérico, que é acrescentado ao conjunto de padrões aprendidos pelo sistema.

Além do casamento dos padrões acima descritos, este trabalho também utiliza o casamento de bigramas: duplas de palavras que sempre ocorrem na mesma ordem. Por não serem generalizáveis, os bigramas possuem um grande valor para a determinação das classes onde ocorrem. Um exemplo é o bigrama "de acordo", que quando iniciando um segmento de uma reivindicação, geralmente determina uma referência (REF).

3.5.2 Extração e codificação dos atributos

Uma vez que os segmentos tenham sido alinhados, é possível obter todos os atributos descritos na Seção 3.4.1. As próximas seções descrevem respectivamente como cada atributo é extraído e como estes são codificados para o mecanismo de aprendizado utilizado, o modelo de WANN WiSARD.

Método de extração

O primeiro passo da extração de atributos é obter a sequência de palavras do segmento alinhado. Isto é feito pela simples leitura em ordem das folhas do nó alinhado na árvore sintática. As folhas também carregam as respectivas classes gramaticais. Em seguida:

1. O vetor de frequência das classes gramaticais é preenchido com a contagem de cada classe no segmento.

2. A posição relativa da primeira ocorrência de cada classe é mantida por um contador, e registrada na posição correspondente do vetor de ordem das classes gramaticais.

3. O número de palavras é registrado no atributo correspondente.

4. Para cada palavra, verifica-se esta está capitalizada (primeira letra maiúscula) ou totalmente escrita em maiúsculas. Caso todas estejam, registra o atributo "formato título" como verdadeiro e do contrário, falso.

5. Registra-se a classe sintática do nó alinhado no atributo correspondente.
6. Registra-se a classe sintática do pai do nó alinhado no atributo correspondente.
7. Registra-se a classe de segmento do pai do nó alinhado na árvore de segmentos, no atributo correspondente.
8. Registra-se a classe do nó anterior na árvore de segmentos (pela visita em profundidade e pré-ordem) no atributo correspondente.

Exemplo: os atributos do segmento "OBJ_PAT" da Figura 3.9 possuem os seguintes valores, considerando a seguinte ordenação para o vetor de classes gramaticais: [A, ART, PNT, P, CONJ, N, CARD, V].

- Freq. das classes gramaticais: [1, 1, 0, 0, 0, 1, 1, 0]
- Ordem das classes gramaticais: [3, 1, 0, 0, 0, 2, 4, 0]
- Número de palavras: 4
- Formato título: falso
- Classe sintática: NP
- Classe sintática so pai: PP
- Classe de segmento pai: CARACT_ASSUNTO
- Classe de segmento anterior: CARACT_ASSUNTO

Binarização

Conforme visto na Seção 2.4.2, a Rede Neural WiSARD utiliza como entrada apenas padrões binários. Portanto é necessário converter os valores dos atributos para uma representação binária adequada, em um processo chamado de *binarização*. Os atributos foram então divididos em dois tipos: numéricos e nominais. Um esquema de binarização foi adotado para cada tipo. Os atributos numéricos são: os vetores de frequência e ordem das classes gramaticais e o número de palavras. O restante são atributos nominais.

Para os atributos numéricos, foi adotado o esquema de binarização *Termômetro*, também conhecido como *Unário*. Neste esquema, os valores são escalonados para um intervalo $[0, K]$, onde K é o número de bits desejado para a representação, e então arredondados para o inteiro mais próximo. Em seguida um vetor com K bits é preenchido com tantos "1"s à direita em quantidade igual ao valor arredondado do atributo, e com zero nas posições restantes. Este esquema tem como característica

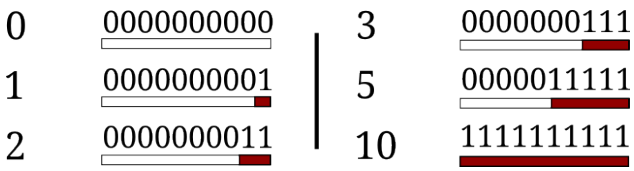


Figura 3.10: Exemplo do esquema de binarização termômetro para um vetor de 10 bits. A proporcionalidade à distância numérica é preservada na distância de Hamming dos valores binarizados. Distâncias maiores implicam em maior contraste entre os valores.

manter a distância de Hamming entre duas representações proporcional a distância numérica entre os valores representados. A Figura 3.10 ilustra este esquema.

Para os atributos nominais, assume-se que não há ordem definida para os valores e que portanto estes são equidistantes. Para manter esta característica, foi escolhido um esquema que maximiza a distância de Hamming entre quaisquer dois pares de valores. Como os vetores de bits resultantes devem ter um tamanho fixo, distancias maiores implicam em um contraste maior entre as representações binárias, tornando-as mais fáceis de distinguir.

Os atributos nominais deste trabalho assumem apenas um pequeno conjunto de valores: 16 classes gramaticais, 24 classes sintáticas e 9 classes de segmento (contando a "não classe"). Isto permitiu que, para cada atributo fosse gerada uma lista de representações binárias equidistantes obtidas a partir de um simples algoritmo de força bruta, uma para cada valor. O algoritmo verifica todas as permutações de um vetor de K bits computando a distância de Hamming para todos os elementos de um conjunto iniciado com "$000\ldots00$". Se a distância for igual a $K/2$, o vetor é acresentado ao conjunto. Este algoritmo gera um conjunto com K vetores binários equidistantes, sendo viável para $K \leq 32$. A distância $K/2$ foi escolhida por ser a maior distância possível para mais de dois valores e menos de K valores, usando um vetor de K bits. No caso de 2 valores, a distância máxima é igual a K (todos os bits 1 ou todos os bits 0). Pode-se aumentar o valor de K e a distância no caso de atributos com até 32 valores distintos. Para isto basta duplicar os vetores (e.g., $0110 \rightarrow 01100110$). Para $K > 32$, o número de combinações a serem verificadas torna-se grande demais e um algoritmo de menor custo computacional seria necessário. O desenvolvimento de métodos eficientes para construção de códigos equidistantes é um tema antigo de estudo entre matemáticos [84], com avanços incrementais ao longo dos anos [85, 86]. Entretanto, a ausência da necessidade imediata levou à opção de não explorar tais métodos no escopo deste trabalho. A Figura 3.11 ilustra um conjunto de vetores binários equidistantes, com maior distância possível.

O valor de K é ajustado para cada atributo, dependendo do número de valores possíveis e da importância que se deseja atribuir em relação aos demais. Valores maiores para K fazem a rede WiSARD armazenar mais bits relativos ao atributo,

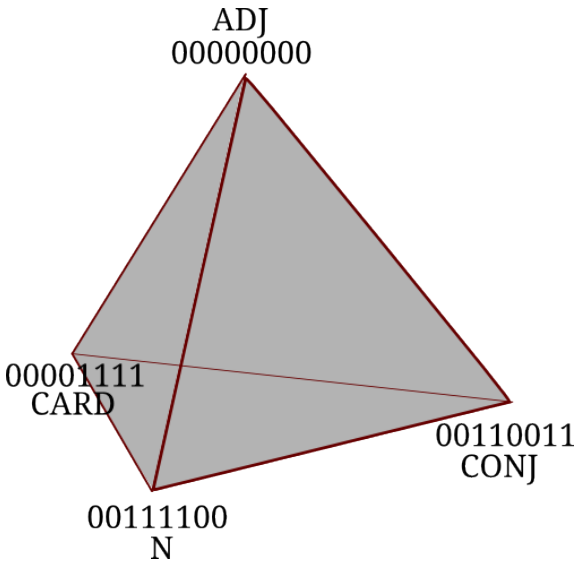


Figura 3.11: Exemplo do esquema de representação binária nominal para um vetor de 8 bits, com cada valor representando uma classe gramatical. A distância de Hamming máxima neste caso é igual a 4. A ausência de ordem ou distância natural entre os valores é preservada ao manter o contraste equivalente entre dois pares quaisquer de representações binárias.

dando a ele um peso maior na classificação.

3.5.3 Configuração da WiSARD

Sendo um modelo de classificador simples, a rede WiSARD precisa de apenas dois parâmetros de ajuste: o tamanho da entrada e o número de linhas de endereçamento, ou tamanho do endereço. Para o primeiro, foram atribuídos os valores de K para cada atributo da seguinte maneira:

- Frequência. das classes gramaticais: 40 para cada posição
- Ordem das classes gramaticais: 20 para cada posição
- Número de palavras: 200
- Formato título: 128
- Classe sintática: 128
- Classe sintática so pai: 128
- Classe de segmento pai: 128
- Classe de segmento anterior: 128

Estes valores de K foram escolhidos após consideração do número de possíveis valores de cada atributo e da importância relativa de cada um. Por exemplo, a frequência máxima admitida para cada classe gramatical é 40. Valores acima do máximo são computados como sendo o máximo. Da mesma forma, o tamanho máximo de sentença admitido pelo sistema é 200 palavras. No caso do vetor de ordem das classes gramaticais, o valor máximo é 10 (10 classes possíveis), mas a distância na codificação termômetro foi duplicada para aumentar a importância relativa desse atributo. O valor de K para os demais atributos também foi obtido pelo ajuste da importância relativa, feito experimentalmente através de testes para medir a capacidade discriminatória do conjunto de atributos, usando algoritmos de classificação conhecidos (Seção 4.1).

A entrada da WiSARD é formada pela concatenação das representações binárias de todos os atributos. Logo, o tamanho do entrada é $40 \times 16 + 20 \times 24 + 200 + 128 \times 5 = 1960$.

Para o número de linhas de endereçamento, foram testadas as potências de 2 entre 1 e 128, todos divisores de 1960. O valor 1 apresentou maior estabilidade nas respostas, mas o pior tempo de resposta, também não tendo o melhor desempenho de classificação, o valor 64 apresentou o melhor tempo de resposta e um relativo bom desempenho de classificação, especialmente quanto à abrangência (recall), enquanto o valor 128 apresentou o pior desempenho de classificação e estabilidade. Valores entre 64 e 1 apresentaram apenas pequenas melhorias de estabilidade e nenhuma melhoria no desempenho de classificação. Desta forma, o número de linhas de endereçamento foi estabelecido experimentalmente em 64.

A rede WiSARD foi configurada para usar o *bleaching* em caso de empate nas respostas de dois ou mais discriminadores. Isto é mais comum para segmentos com características similares, mas que possuem detalhes diferentes, como números referenciando ilustrações ou outras reivindicações no documento (classes ILUST_REF e REF_REIVIND_NUM).

3.6 Extração de segmentos semânticos

A fase de extração possui um fluxo mais simples, pois há muito menos informações disponíveis. Estas devem ser acrescentadas aos dados observados à partir daquilo que foi aprendido pelo sistema na fase de treinamento, na forma de hipóteses sobre os atributos dos possíveis segmentos.

O primeiro passo da extração é a obtenção dos trechos da reivindicação apresentada que podem ser considerados como segmentos. Estes são chamados *candidatos a segmento*. Para isso, é feita a análise sintática da reivindicação e a árvore de constituintes resultante é percorrida em profundidade e pré-ordem, preenchendo uma

lista com todos os nós sintáticos da sentença. Em seguida, o conjunto de pares sintático-semânticos é utilizado para determinar quais nós sintáticos nunca ocorrem em segmentos, i.e., nunca são alinhados. Estes são excluídos da lista, restando apenas os que podem formar pares conhecidos. A lista resultante contém os candidatos a segmento que serão processados no restante desta fase.

O próximo passo é preencher os atributos de cada candidato, usando o mesmo método apresentado na Seção 3.5.2, com a diferença de que não há árvore de segmentos, então todas as operações são feitas como se o candidato fosse um segmento. Os únicos atributos que não podem ser preenchidos desta forma são as classes de segmento pai e anterior. Estes serão preenchidos conforme o andamento do processo de extração.

A partir deste ponto, o sistema inicia um ciclo de hipótese e validação para cada candidato, que funciona confome descrito nos Algoritmos 2, 3, 4, 5, inspirados no trabalho de De Gregorio et al. [87].

Algoritmo 2 Algoritmo de hipótese e validação de candidatos a segmento semântico – Ciclo principal.

```
1: usando lista_de_candidatos
2: enquanto houver candidatos não decididos faça
3:     formular_hipoteses(lista_de_candidatos)
4:     validar_hipoteses(lista_de_candidatos)
5: fim enquanto
```

Algoritmo 3 Algoritmo de hipótese e validação de candidatos a segmento semântico – Formulação de hipóteses.

```
 1: procedure FORMULAR_HIPOTESES(lista_de_candidatos)
 2:     para cada candidato não decidido em lista_de_candidatos faça
 3:         respostas_wisard ← lista_vazia()
 4:         para cada candidato não marcado em lista_de_candidatos faça
 5:             preenche_classe_pai(candidato)
 6:             preenche_classe_anterior(candidato)
 7:             respostas_wisard.adiciona(classifica_com_wisard(candidato))
 8:         fim para
 9:         candidato_selecionado ← seleciona_maior_resposta(respostas_wisard)
10:         candidato_selecionado.classe ← classe_da_maior_resposta(respostas_wisard)
11:         candidato.marcado ← Verdadeiro
12:         se (candidato.classe em candidato.classes_rejeitadas) então ▷
    candidato não é segmento
13:             candidato.marcado ← Falso
14:             candidato.decidido ← Verdadeiro
15:         fim se
16:     fim para
17: fim procedure
```

Algoritmo 4 Algoritmo de hipótese e validação de candidatos a segmento semântico – Validação de hipóteses.

```
1: procedure VALIDAR_HIPOTESES(lista_de_candidatos)
2:     para cada candidato marcado e não decidido em lista_de_candidatos faça
3:         se padrão morfológico do candidato casa com algum de candidato.classe então
4:             prossegue com padrão morfológico OK.
5:         senão
6:             prossegue com padrão morfológico NÃO OK.
7:         fim se
8:         se bigrama iniciando o candidato casa com algum de candidato.classe então
9:             a hipótese está OK.
10:            candidato.decidido ← Verdadeiro
11:        senão
12:            se padrão morfológico NÃO OK então
13:                candidato.marcado ← Falso
14:            fim se
15:        fim se
16:        se par (candidato.classe, candidato.classe_pai) está nos pares hierárquicos então
17:            candidato.marcado ← Falso
18:            candidato.classe_pai ← Vazio
19:        fim se
20:        se par (candidato.classe, candidato.classe_sintática) está nos pares semântico-sintáticos
   então
21:            candidato.marcado ← Falso
22:        fim se
23:        se posição do candidato em relação aos marcados está no conjunto de ordem para candidato.classe
   então
24:            candidato.marcado ← Falso
25:        fim se
26:        se número de palavras do candidato está dentro dos limites do conjunto de tamanhos
   para candidato.classe então
27:            candidato.marcado ← Falso
28:        fim se
29:        se candidato.marcado = Verdadeiro então
30:            candidato.decidido ← Verdadeiro
31:        senão
32:            candidato.classes_rejeitadas.adiciona(candidato.classe)
33:            candidato.marcado ← Falso
34:        fim se
35:    fim para
36: fim procedure
```

A cada iteração do algoritmo, os candidatos desmarcados têm suas classes atribuídas por hipótese apagadas. Se havia filhos atribuídos ao candidato, por meio do *preenche_ classe_ pai*, estes filhos têm a classe pai apagada, para que suas hipóteses

Algoritmo 5 Algoritmo de hipótese e validação de candidatos a segmento semântico – Preenchimento de atributos.

```
 1: procedure PREENCHE_CLASSE_PAI(candidato, lista_de_candidatos)
 2:     para cada outro_candidato marcado em lista_de_candidatos faça
 3:         se candidato contido em outro_candidato então
 4:             candidato.classe_pai ← outro_candidato.classe
 5:         fim se
 6:     fim para
 7: fim procedure
 8: procedure PREENCHE_CLASSE_ANTERIOR(candidato, lista_de_candidatos)
 9:     se ultimo_candidato_marcado é anterior a candidato na sentença então
10:         candidato.classe_anterior ← ultimo_candidato_marcado.classe
11:     fim se
12: fim procedure
```

possam também ser corrigidas.

Ao fim de todo o processo, os candidatos desmarcados são descartados e o restante é colocado em uma lista, na ordem em que aparecem na sentença.

3.7 Extração de relações

3.7.1 Aplicação das regras de relacionamento

Considerando a informação já contida nos segmentos semânticos, é possível elaborar regras simples para relacionar os casos mais comuns de relações entre os segmentos. Esta é uma informação que faz parte do modelo de representação do conhecimento, junto com o modelo de segmentação semântica, sendo manualmente codificada.

Para fins de automatização, foi utilizado um esquema de notação similar a expressões regulares para codificar as regras. Estas representam um padrão sequencial simples a ser encontrado na lista de segmentos. As regras utilizadas neste trabalho são apresentadas na Tabela A.2, Apêndice A.

As regras são aplicadas da seguinte forma: a lista de segmentos extraídos é lida em ordem, e para cada segmento é verificada a presença de um dos padrões declarados nas regras. Caso um padrão seja identificado, o resultado esperado da regra é produzido e o processo continua até que a lista de segmentos chegue ao fim. Colisões de regras são possíveis, resultando na produção de dois ou mais resultados diferentes para o mesmo padrão.

Cada regra produz uma tripla *(sujeito, predicado, objeto)* que, ao ser combinada com outras, forma o grafo de relações semânticas da reivindicação. Os grafos de todas as reivindicações de um documento podem ser combinados da mesma forma.

A Figura 3.12 ilustra o grafo de relações que deve ser extraído a partir da reivindicação exemplo da Figura 3.8, usando as regras apresentadas. A união de todos os grafos de um documento permite também a construção do grafo de reivindicações

(Seção 2.1.2) para a patente.

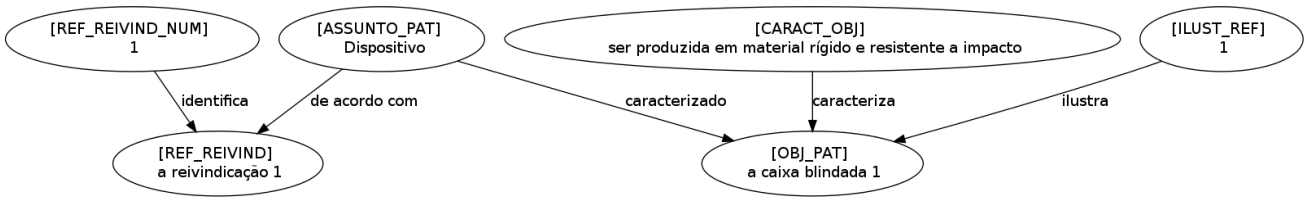


Figura 3.12: Exemplo de grafo de relações para uma reivindicação. É possível identificar facilmente o tópico e referências usadas na reivindicação.

3.7.2 Instanciação das ontologias

Com o grafo de relações pronto, pode-se instanciar entidades de uma ontologia de patentes através do alinhamento das entidades do grafo com as da ontologia alvo.

As entidades do grafo podem ser identificadas através de um sistema de Resolução de Entidades Nomeadas (NER, Seção 2.3.3), como os participantes da avaliação conjunta HAREM [88]. Entretanto, Durante a elaboração deste trabalho não foi encontrado um sistema de NER que mapeasse entidades em uma ontologia de patentes em português. Uma alternativa é usar um sistema de busca de termos que utilize algoritmos de mapeamento ontológico ou algoritmos de similaridade semântica. Para a língua inglesa, um exemplo da primeira categoria é o DBPedia Spotlight [89] que busca entidades da DBpedia [22], e da segunda categoria, o EasyESA [90], que usa os textos da Wikipedia como base de conhecimento e a Wordnet para desambiguação. Também não foram encontrados sistemas com estas características voltados ao português durante a elaboração deste trabalho.

O alinhamento simples dos termos com uma wordnet também é possível e interessante, especialmente para os objetos de patente, já que permite a desambiguação de certos termos e facilita o mapeamento de entidades em documentos diferentes, inclusive em outras línguas. A Figura 3.13 mostra o resultado da busca pelo termo "blindagem" na OpenWordNet-PT [91]. A Figura 3.14 ilustra o alinhamento com reivindicações de documentos diferentes e a visão de objetos similares. Este mesmo processo pode ser usado para realizar a resolução de correferência (Seção 2.3.4), ao unir segmentos anafóricos[4] de um mesmo grafo.

[4]Aqueles que fazem referência a segmentos anteriores no texto

- S: (n) testa, armadura, concha, **blindagem**, couraça, coiraça (revestimento de aço com que se protegem os navios encouraçados contra a artilharia. revestimento externo da casca de sementes. revestimento de navios com ferro ou outro metal.)
- S: (n) revestimento, **blindagem** (material que reveste um objeto, conferindo-lhe proteção contra projétil ou elemento potencialmente danoso.)

Figura 3.13: Resultado da busca da palavra "blingagem" na OpenWordNet-PT. Os diferentes significados retornados pela busca (um em cada item da lista) podem ser utilizados para desambiguar termos usados na patente, visto que geralmente são compostos por mais de uma palavra. O termo "blindagem protetora" seria desambiguado na segunda entrada da lista.

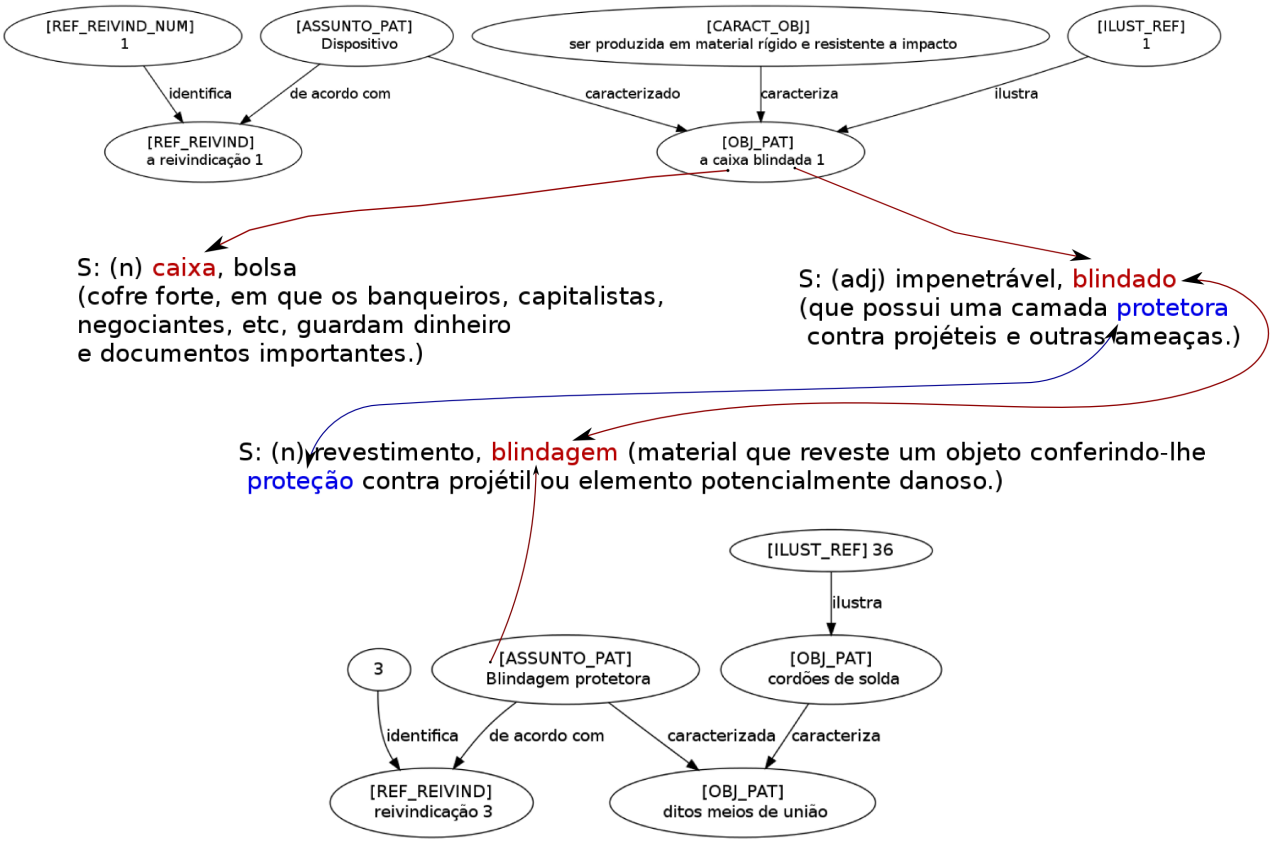


Figura 3.14: Exemplo de alinhamento de duas reivindicações de documentos de patente distintos com a OpenWordNet-PT. Documentos diferentes podem ser comparados semanticamente quanto à proximidade de conceitos abordados, particularmente nos tópicos e objetos declarados.

Capítulo 4

Ambiente Experimental e Resultados

4.1 Escolha dos atributos e avaliação do potencial discriminatório

Conforme visto na Seção 3.4.1, os atributos usados na classificação dos segmentos foram selecionados com base em critérios linguísticos e experimentais, aplicados a textos de natureza declarativa. Uma vez selecionados, era importante verificar se os atributos realmente serviriam para distinguir um tipo de segmento de outro antes de prosseguir com a construção do sistema. O potencial discriminatório do conjunto de atributos foi avaliado através do seguinte procedimento:

1. Realização do alinhamento sintático-semântico e preencher valores dos atributos, conforme descrito na Seção 3.5.2;
2. Geração de uma tabela com uma linha para cada segmento alinhado e uma coluna para cada atributo, mais uma coluna para a classe do segmento;
3. Teste da tabela pelo método *10-fold cross validation*, usando um algoritmo de classificação conhecido;
4. Conferência dos resultados do teste.

O algoritmo de classificação escolhido para o teste das tabelas foi o *Perceptron Multicamada* [59], uma ANN tradicional (ver Seção 2.4.1). A escolha de um algoritmo classificador diferente da WiSARD foi feita para verificar deficiências de implementação e otimização da WiSARD em uma etapa posterior da construção do sistema.

O conjunto inicial de atributos não incluía "formato título" e "classe do segmento anterior". A Tabela 4.1 mostra os resultados do teste de classificação para este conjunto, usando as medidas típicas de acurácia e abrangência no conjunto. Após

melhor observação, os atributos citados foram incluídos. A Tabela 4.2 mostra os resultados com todos os atributos.

Também foi levantada a hipótese de algum atributo estar atrapalhando na classificação. O teste então foi repetido eliminando cada um dos atributos, resultando sempre em pior desempenho. Assim foi definido o conjunto final de atributos.

Tabela 4.1: Resultados do teste de classificação para os atributos do modelo de segmentação, excluindo "formato título" e "classe do segmento anterior". O teste mede o potencial discriminatório dos atributos, i.e., a capacidade dos atributos de servir à diferenciação entre as diferentes classes.

Classe	Acurácia	Abrangência	Medida F1
ASSUNTO_PAT	0.82	0.86	0.84
OBJ_PAT	1.0	1.0	1.0
ILUST_REF	0.97	1.0	0.98
CARACT_ASSUNTO	0.70	0.63	0.66
CARACT_OBJ	1.0	0.92	0.96
REF_REIVIND	1.0	1.0	1.0
REF_REIVIND_NUM	1.0	1.0	1.0
Total	0.96	0.96	0.96

Tabela 4.2: Resultados do teste de classificação para os atributos do modelo de segmentação, incluindo todos os atributos.

Classe	Acurácia	Abrangência	Medida F1
ASSUNTO_PAT	0.96	1.0	0.98
OBJ_PAT	1.0	1.0	1.0
ILUST_REF	0.98	1.0	0.99
CARACT_ASSUNTO	1.0	0.91	0.95
CARACT_OBJ	1.0	0.96	0.98
REF_REIVIND	1.0	1.0	1.0
REF_REIVIND_NUM	1.0	1.0	1.0
Total	0.99	0.99	0.99

Pode ser notado nas tabelas 4.1 e 4.2 que a classe *REF* não está presente. Isto se deve ao fato de que o algoritmo de alinhamento sintático-semântico não conseguiu alinhar nenhuma instância de *REF* no conjunto de reivindicações analisado. O motivo disso são os erros provenientes da análise sintática, que separam partes adjacentes de uma mesma frase em nós diferentes da árvore de constituintes. No caso do LX-Parser, expressões como "de acordo com" são especialmente vulneráveis a esse tipo de erro. Esse problema é minimizado com o uso dos padrões morfológicos (Seção 3.5.1), que conseguem capturar as expressões mais frequentes para cada tipo de segmento.

O teste para avaliação do potencial discriminatório foi repetido posteriormente, em duas condições diferentes: (*i*) mudando o algoritmo de classificação e (*ii*) removendo ambos os atributos semânticos (classe do segmento pai e classe do segmento anterior). O objetivo da condição (*i*) foi averiguar o desempenho de outros algoritmos em termos de acurácia e abrangência, especialmente a rede WiSARD, usada no sistema AS^{2}ABER. A condição (*ii*) foi testada para analisar o comportamento dos classificadores em uma situação mais próxima daquela encontrada durante a operação do sistema AS^{2}ABER, onde tipicamente pelo menos um dos dois atributos estará faltando ou incorreto. A tabela 4.3 mostra os resultados do teste de classificação usando o Perceptron Multicamada para a condição (*ii*).

Tabela 4.3: Resultados do teste de classificação com o algoritmo Perceptron Multicamada para os atributos do modelo de segmentação, excluindo os atributos semânticos.

Classe	Acurácia	Abrangência	Medida F1
ASSUNTO_PAT	0.62	0.68	0.65
OBJ_PAT	0.78	0.77	0.77
ILUST_REF	0.84	0.97	0.90
CARACT_ASSUNTO	0.20	0.09	0.13
CARACT_OBJ	0.57	0.48	0.52
REF_REIVIND	0.5	0.71	0.59
REF_REIVIND_NUM	0.33	0.15	0.21
Total	0.70	0.73	0.71

Além do Perceptron Multicamada, os algoritmos de classificação testados foram:

- *C4.5* [92] (árvores de decisão).
- *RIPPER* [93] (regras proposicionais).
- *SVM* [94] (com kernel linear).
- *Naive Bayes* [95].
- *WiSARD* [65].

As tabelas 4.4 a 4.8 mostram o resultado dos testes em ambas as condições (*i*) e (*ii*) para cada algoritmo.

Tabela 4.4: Resultados do teste de classificação com o algoritmo C4.5 para os atributos do modelo de segmentação, incluindo e excluindo os atributos semânticos respectivamente.

Classe	Acurácia		Abrangência		Medida F1	
	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.
ASSUNTO_PAT	0.87	0.86	0.91	0.82	0.89	0.84
OBJ_PAT	1.00	0.77	0.97	0.89	0.98	0.83
ILUST_REF	0.96	0.81	1.00	0.99	0.98	0.89
CARACT_ASSUNTO	0.80	0.50	0.73	0.36	0.76	0.42
CARACT_OBJ	1.00	0.60	0.96	0.33	0.98	0.43
REF_REIVIND	1.00	0.60	1.0	0.43	1.0	0.50
REF_REIVIND_NUM	1.00	0.50	1.0	0.15	1.0	0.23
Total	0.97	0.74	0.97	0.77	0.97	0.74

Tabela 4.5: Resultados do teste de classificação com o algoritmo ripper para os atributos do modelo de segmentação, incluindo e excluindo os atributos semânticos respectivamente.

Classe	Acurácia		Abrangência		Medida F1	
	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.
ASSUNTO_PAT	0.92	0.86	1.00	0.82	0.96	0.84
OBJ_PAT	0.95	0.73	0.99	0.88	0.97	0.80
ILUST_REF	0.99	0.84	0.99	0.94	0.99	0.89
CARACT_ASSUNTO	1.00	0.33	0.82	0.18	0.90	0.24
CARACT_OBJ	0.96	0.52	0.82	0.44	0.88	0.48
REF_REIVIND	1.00	0.82	1.00	0.64	1.00	0.72
REF_REIVIND_NUM	1.00	0.33	1.00	0.05	1.00	0.09
Total	0.97	0.72	0.97	0.75	0.97	0.72

Tabela 4.6: Resultados do teste de classificação com o algoritmo SVM para os atributos do modelo de segmentação, incluindo e excluindo os atributos semânticos respectivamente.

Classe	Acurácia		Abrangência		Medida F1	
	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.
ASSUNTO_PAT	0.92	0.95	1.00	0.86	0.96	0.90
OBJ_PAT	0.98	0.81	1.00	0.81	0.99	0.81
ILUST_REF	0.98	0.82	1.00	0.97	0.99	0.89
CARACT_ASSUNTO	1.00	0.18	0.73	0.18	0.84	0.18
CARACT_OBJ	1.00	0.65	0.93	0.56	0.96	0.60
REF_REIVIND	1.00	0.53	0.93	0.64	0.96	0.58
REF_REIVIND_NUM	1.00	0.43	1.00	0.15	1.00	0.22
Total	0.98	0.75	0.98	0.76	0.98	0.75

Tabela 4.7: Resultados do teste de classificação com o algoritmo Naive Bayes para os atributos do modelo de segmentação, incluindo e excluindo os atributos semânticos respectivamente.

Classe	Acurácia		Abrangência		Medida F1	
	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.
ASSUNTO_PAT	0.95	0.95	0.90	0.82	0.93	0.88
OBJ_PAT	0.86	0.84	0.80	0.75	0.83	0.79
ILUST_REF	1.00	0.84	0.95	0.97	0.97	0.90
CARACT_ASSUNTO	0.60	0.60	0.27	0.27	0.37	0.37
CARACT_OBJ	0.63	0.62	0.70	0.67	0.67	0.64
REF_REIVIND	0.54	0.48	0.93	0.93	0.68	0.63
REF_REIVIND_NUM	0.80	0.30	1.00	0.15	0.89	0.2
Total	0.86	0.76	0.85	0.77	0.85	0.75

Tabela 4.8: Resultados do teste de classificação com o algoritmo WiSARD para os atributos do modelo de segmentação, incluindo e excluindo os atributos semânticos respectivamente.

Classe	Acurácia		Abrangência		Medida F1	
	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.	c/ atribs.	s/ atribs.
ASSUNTO_PAT	0.50	0.48	0.95	0.75	0.66	0.58
OBJ_PAT	0.95	0.71	0.93	0.73	0.94	0.72
ILUST_REF	0.88	0.85	0.98	0.13	0.93	0.23
CARACT_ASSUNTO	0.00	0.00	0.00	0.00	0.00	0.00
CARACT_OBJ	1.00	0.64	0.36	0.36	0.53	0.46
REF_REIVIND	0.86	0.6	0.92	0.92	0.89	0.73
REF_REIVIND_NUM	0.89	0.18	0.89	0.83	0.89	0.30
Total	0.85	0.66	0.85	0.49	0.85	0.56

Os algoritmos Perceptron Multicamada, C4.5, RIPPER, SVM e *Naive Bayes* foram executados através da ferramenta WEKA [96]. A implementação de SVM usada foi a LibSVM [97]. Para WiSARD foi usada uma implementação própria, escrita em linguagem *Python*, e sua configuração está descrita na Seção 3.5.2. As configurações usadas para os demais algoritmos foram as definições padrão da ferramenta WEKA. Eles são apresentados no Apêndice B.

Como pode ser observado nas Tabelas 4.3 a 4.8, os atributos semânticos possuem um grande impacto no desempenho de classificação. Quando comparadas com a Tabela 4.1, é possível notar a importância maior do atributo "segmento pai", reforçando o caráter hierárquico do modelo de informações adotado. Também pode-se notar que o desempenho de todos os algoritmos é afetado de forma similar pela ausência dos atributos semânticos, embora classes específicas possam se comportar de maneira diferente dependendo do algoritmo usado.

O desempenho inferior do classificador WiSARD, especialmente na ausência dos atributos semânticos, pode ser atribuído à falta de otimização na codificação dos atributos para binário. Conforme explicado na Seção 2.4.2, a binarização é um fator crítico no desempenho da rede e deve ser feita para cada atributo. O esquema de binarização usado neste trabalho (Seção 3.5.2) faz o ajuste do tamanho da representação binária para cada atributo em função de sua importância relativa, obtida empiricamente. Entretanto, foi notado em análise posterior aos experimentos que para os atributos vetoriais (vetor de *POS-tags*, vetor de ordens gramaticais) há uma considerável diferença entre a importância relativa de cada elemento do vetor. Esta diferença foi corretamente capturada pelos demais algoritmos de classificação testados. Nesse caso, a binarização deve ser modificada para levar em conta esse fator, considerando cada elemento dos vetores como um atributo diferente. Os estudos feitos sobre o uso do classificador WiSARD em um problema com grande quantidade de atributos [70] indicam que o desempenho deste classificador será equivalente ao do classificador SVM após a otimização da codificação dos atributos.

A escolha do classificador WiSARD é justificada pela relativa simplicidade de implementação em diferentes ambientes computacionais e escalabilidade [70], e também pela possibilidade de adicionar informação sobre novas reivindicações conforme estas são obtidas, i.e., treinamento *online*. A capacidade de treinamento *online* é de grande valor para este trabalho, pois viabiliza a análise incremental de grandes quantidades de documentos.

Durante a execução dos testes, foi notado que os algoritmos C4.5 e RIPPER foram muito mais rápidos (> 20 vezes) do que os demais. Entretanto, em uma análise intuitiva da saída, algumas das regras geradas por estes dois algoritmos apresentariam problemas de generalização caso fossem aplicados a um conjunto mais abrangente de sentenças. Para confirmar isto, seria necessário um corpus maior, o qual ainda não estava disponível. O tempo de execução dos demais algoritmos não foi comparado em função da grande diferença em termos de implementação, o que faria que os resultados refletissem o ambiente de execução, e.g., interpretado vs. nativo, e não os algoritmos em si.

4.2 Avaliação de qualidade da extração

A qualidade da extração dos segmentos pode ser analisada de diferentes pontos de vista, dependendo do uso pretendido para o resultado. Dois pontos de vista foram explorados para a avaliação: (*i*) corretude e abrangência dos segmentos extraídos e (*ii*) qualidade da informação obtida do grafo de relações semânticas. Para cada um foi escolhida uma medida correspondente, que foi analisada nos experimentos.

Para medir a corretude e abrangência da extração, foi escolhida a métrica de

acurácia e abrangência (*precision e recall*) tipicamente usada para avaliar o desempenho de sistemas de classificação. A acurácia mede a proporção dos segmentos extraídos que estão corretos, dados os segmentos conhecidos em um conjunto de reivindicações de teste. A abrangência mede a proporção dos segmentos conhecidos que foi extraída no mesmo conjunto de reivindicações de teste. Entretanto, por este se tratar de um sistema de extração e não apenas de classificação, as duas métricas são calculadas com alguns ajustes, da seguinte maneira:

$$acurácia_{classe} = \frac{\text{Nº de acertos}}{\text{Nº de extrações}} \tag{4.1}$$

$$abrangência_{classe} = \frac{\text{Nº de acertos}}{\text{Nº de segmentos}} \tag{4.2}$$

Os resultados totais são ponderados por classe da seguinte maneira:

$$acurácia_{total} = \frac{\sum_{\forall classe} \frac{\text{Nº de acertos}}{\text{Nº de extrações}} \times \acute{m}inimo(\text{Nº de segmentos, Nº de extrações})}{\sum_{\forall classe} \acute{m}inimo(\text{Nº de segmentos, Nº de extrações})} \tag{4.3}$$

$$abrangência_{total} = \sum_{\forall} \frac{\text{Nº de acertos}}{\text{Nº de segmentos}} \tag{4.4}$$

Um acerto é computado quando o segmento extraído possui uma sobreposição de caracteres de mais de 75% com o segmento correspondente no conjunto verdade. Isto é feito para considerar segmentos com artigos ou preposições em falta ou excesso nas extremidades.
O termo $\acute{m}inimo$(Nº de segmentos, Nº de extrações) é usado para não computar na acurácia extrações que não foram realizados para uma classe, estes contando apenas para a abrangência.

A medida F_1, definida como a média harmônica da acurácia e abrangência, foi incluida para completar a visão dos resultados do sistema.

$$F_1 = 2 \times \frac{acurácia \times abrangência}{acurácia + abrangência} \tag{4.5}$$

Para medir a qualidade da informação obtida, os grafos resultantes da extração de relações foram analisados quanto à possibilidade de recriar toda ou parcialmente a informação contida na reivindicação. Os grafos podem ser usados como uma forma de resumo do conteúdo das patentes, especialmente ao unir os grafos de todas as reivindicações de um documento. Para este fim, em muitos casos o tópico e alguma caracterização já são suficientes para saber a relevância do documento em uma busca.

A presença destes elementos (ASSUNTO_PAT, CARACT_ASSUNTO + OBJ_-PAT) foi utilizada para determinar se um grafo era ou não informativo. O número de grafos informativos foi usado como medida de qualidade do sistema.

O uso da acurácia e abrangência foi feito com a intenção de medir a robustez do sistema do ponto de vista da Recuperação de Informação, mas não visando a comparação com outros sistemas de extração de informação de patentes como [7] e [3], uma vez que não há um consenso sobre o tipo e a forma das informações a serem extraídas. A comparação com outros sistemas de extração de relações semânticas também não foi possível, dado que os arcabouços e *padrões-ouro* necessários para tal são focados em tarefas específicas, como extração de relações parte-todo e de papéis semânticos, que não são o escopo deste trabalho.

O uso da análise de qualidade dos grafos foi feito com o objetivo de avaliar a utilidade do sistema em um cenário de uso simples, cuja premissa era a redução da necessidade de ler os documentos de patente para filtrar a informação desejada.

Além destes, também foi medido o tempo necessário para extração dos segmentos em cada reivindicação, com o objetivo de analisar a possibilidade de operação do sistema em larga escala.

4.3 Experimentos

4.3.1 Amostragem dos documentos

Os documentos usados nos experimentos foram obtidos do sistema de consulta pública do INPI, realizando buscas por 6 assuntos gerais: Eletrônicos, Alimentos, Utilidades, Indústria, Agricultura e Saúde. Para cada assunto, algumas patentes foram escolhidas manualmente, tendo como único critério a presença de uma versão completa e legível do documento em formato PDF. Muitos documentos contam apenas com o resumo da patente e alguns apresentam uma qualidade ruim de impressão ou digitalizaçao, que dificulta a leitura. Um conjunto de documentos foi obtido para cada assunto após as tentativas feitas sobre os resultados da busca, que podiam ser feitas em um número limitado por intervalo de tempo. A escolha dos assuntos foi feita visando testar o sistema com termos e construções frasais que refletissem a variedade textual encontrada nas patentes brasileiras.

Devido ao tempo necessário para aquisição, OCR, limpeza e anotação manual dos documentos, apenas um conjunto pequeno de 20 patentes foi utilizado nos experimentos, em função do tempo necessário para conclusão do trabalho. Destes foram obtidas e anotadas 50 reivindicações, totalizando aproximadamente 400 segmentos. O Apêndice D contém as folhas de rosto das patentes utilizadas nas quais foram gerados grafos informativos, e para cada uma, contém um exemplo de reivindica-

ção selecionada em sua forma original e anotada, junto com o grafo correspondente extraído.

4.3.2 Organização dos experimentos

Os experimentos de acurácia e abrangência foram organizados em 3 tipos, variando o conjunto de reivindicações separadas para treinamento e teste do sistema. A primeira usou 10% das reivindicações para treino e 90% para teste em 10 rodadas (*10-fold cross validation*), a segunda usou 20% para treino e 80% para teste em 5 rodadas (*5-fold cross validation*) e a última usou todas as reivindicações para treino exceto uma, em 49 rodadas (*leave one out*). Nos experimentos *10-fold* e *5-fold*, as reivindicações participantes do treinamento e teste são escolhidas aleatoriamente. No experimento *leave one out*, cada reivindicação é testada uma vez. Cada experimento foi repetido três vezes e a média dos resultados foi utilizada na avaliação.

Os experimentos de qualidade dos grafos foram feitos aplicando as regras de extração de relações às saídas obtidas do *leave one out*, e analisando cada grafo resultante.

4.3.3 Ambiente de execução

Os experimentos foram realizados em um computador com processador Intel Core2 Quad Q6600 (2.4 GHz), 2 GB de memória RAM, e utilizando o sistema operacional Linux Debian "Jessie" (kernel 3.14-1) 64 bits.

Versões de software utilizadas:

- Python 2.7: ambiente de execução do sistema principal.
- Java 1.7: ambiente de execução para o mWANN-Tagger e o LX-Parser.
- NLTK 3.0: usado pelo sistema para leitura as árvores no formato Penn-Treebank.

4.4 Resultados

As Tabelas 4.9, 4.10 e 4.11 mostram os resultados obtidos para os três tipos de experimentos de acurácia e abrangência.

Das 50 reivindicações analisadas, 29 (58%) produziram grafos informativos. As Figuras 4.1 a 4.3 ilustram grafos obtidos.

Como pode ser observado, à medida que mais exemplos vão sendo apresentados ao sistema, sua abrangência melhora pois mais estruturas são aprendidas, contudo

Tabela 4.9: Resultados do teste *10-fold cross validation*.

Classe	Acurácia	Abrangência	Medida F1
ASSUNTO_PAT	0.79	0.47	0.59
OBJ_PAT	0.34	0.05	0.09
ILUST_REF	0.62	0.07	0.12
CARACT_ASSUNTO	0,82	0.45	0.58
CARACT_OBJ	0.36	0.06	0.10
REF	1.00	0.80	0.89
REF_REIVIND	0.92	0.80	0.85
REF_REIVIND_NUM	0.42	0.59	0.49
Total	0.66	0.22	0.33

Tabela 4.10: Resultados do teste *5-fold cross validation*.

Classe	Acurácia	Abrangência	Medida F1
ASSUNTO_PAT	0.77	0.71	0.74
OBJ_PAT	0.31	0.07	0.11
ILUST_REF	0.64	0.07	0.13
CARACT_ASSUNTO	0.74	0.52	0.61
CARACT_OBJ	0.34	0.12	0.18
REF	1.00	0.89	0.94
REF_REIVIND	0.89	0.89	0.89
REF_REIVIND_NUM	0.40	0.73	0.52
Total	0.62	0.26	0.37

Tabela 4.11: Resultados do teste *leave one out*

Classe	Acurácia	Abrangência	Medida F1
ASSUNTO_PAT	0.69	0.93	0.79
OBJ_PAT	0.29	0.40	0.33
ILUST_REF	0.96	0.39	0.55
CARACT_ASSUNTO	0.50	0.67	0.58
CARACT_OBJ	0.49	0.39	0.43
REF	1.0	1.0	1.0
REF_REIVIND	0.89	0.89	0.89
REF_REIVIND_NUM	0.67	0.77	0.71
Total	0.56	0.51	0.53

mais ruído também é aprendido, reduzindo a acurácia, embora não na mesma proporção. A acurácia e abrangência de classes importantes para o problema, como ASSUNTO_PAT e CARACT_ASSUNTO, indica que os resultados do sistema já podem ser usados para auxiliar na sumarização dos documentos de patente. A medida de qualidade dos grafos extraídos corrobora com esta visão, principalmente quando considerado o fato de que os grafos não informativos podem ter alta acurácia, embora com baixa abrangência. Grafos não informativos podem complementar

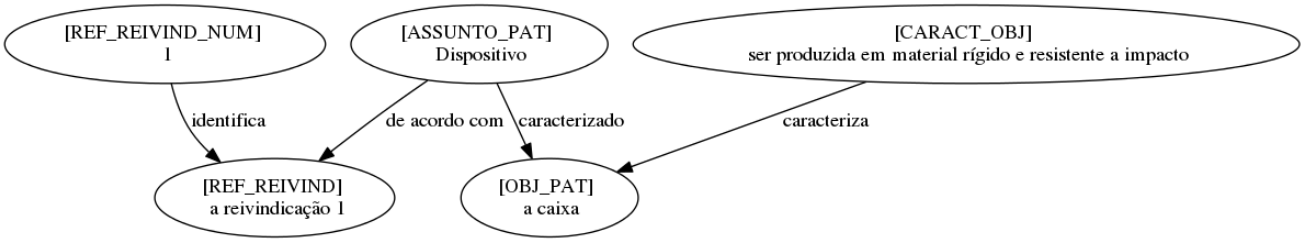


Figura 4.1: Exemplo de grafo informativo obtido do sistema. Quando comparado com a Figura 3.12, é possível observar a ausência da referência à figura, no texto extraído e nó correspondente. O assunto, um objeto que o caracteriza e detalhes sobre esse objeto estão presentes.

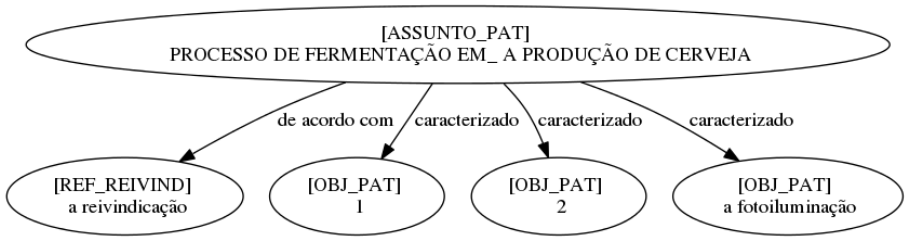


Figura 4.2: Exemplo de grafo informativo obtido do sistema. "1" e "2" foram classificados de forma errada como objetos da patente, quando são na verdade referências a reivindicações. A reivindicação referenciada ficou sem identificador. O terceiro objeto está correto e caracteriza o assunto.

Figura 4.3: Exemplo de grafo não informativo obtido do sistema. O assunto está incompleto e "2" foi classificado de forma errada como objeto da patente, quando é na verdade uma referência a reivindicação. A caracterização do assunto não é possível.

a informação de um documento quando unidos aos grafos das demais reivindicações.

Algumas classes possuem uma abrangência muito baixa. Estas classes são aquelas cuja variedade de estruturas é alta (e.g., OBJ_PAT, CARACT_OBJ) e as que estão tipicamente atreladas a estas (e.g., ILUST_REF). Esse fato, junto à grande distância entre os testes de potencial discriminatório e os testes de extração indicam uma deficiência no processo de classificação, quando os atributos estão incompletos ou contendo erros. Seguindo o mesmo raciocínio, as classes com maior regularidade apresentam abrangência alta (e.g., ASSUNTO_PAT). O tempo médio de processamento por sentença foi de aproximadamente 8 segundos, o que não atende ao

processamento de documentos em larga escala. Entretanto, aproximadamente 75% desse tempo é gasto na formulação de hipóteses, sendo a principal causa do consumo de tempo a eliminação repetida de hipóteses erradas. Desta forma, melhorar as hipóteses traria um ganho significativo de desempenho. O uso de código interpretado (Python) também contribuiu para o alto tempo de resposta.

Classes com extrema regularidade (e.g., REF) são quase totalmente tratadas via padrões morfológicos e portanto dependem muito pouco do classificador. Isso também explica os resultados das classes atreladas (e.g., REF_REIVIND).

Dessa forma, melhorias aplicadas ao classificador devem aumentar a acurácia das classes menos regulares e com isso melhorar muito a abrangência das classes atreladas. A otimização da codificação dos atributos é uma melhoria possível. Uma outra possibilidade é a divisão do classificador em múltiplas instâncias, cada uma orientada a um atributo, visando diminuir a influência dos erros de análise sintática. O algoritmo de hipótese e validação também pode ser modificado para atuar com seleção prévia de classes a serem testadas, evitando a formulação de hipóteses erradas. Alguns testes feitos com uma versão pouco refinada desta abordagem mostraram ganhos substanciais de acurácia e abrangência, mas foram incapazes de analisar todo o conjunto de reivindicações. Mecanismos de melhoria das árvores de constituintes também podem ser considerados. O tempo de resposta do sistema ainda é alto, mas há ainda muito espaço para otimização.

Como consideração importante de melhoria também está a adoção de heurísticas específicas para o tratamento de reivindicações, como a estratégia para quebra de sentenças apresentada em [3]. Por utilizar um modelo de processamento de texto genérico, este trabalho não se beneficia de regras específicas para patentes. Por fim, e não menos importante, está a necessidade de uma massa maior de documentos para experimentação, que permitirá uma avaliação melhor do sistema quando exposto a uma variedade ainda maior de assuntos e construções frasais.

Capítulo 5

Conclusões

5.1 Considerações finais

A extração de informações em patentes apresenta um conjunto de desafios que englobam uma parcela considerável das tarefas conhecidas de Processamento de Linguagem Natural, acrescentando seu próprio conjunto de dificuldades. Entre estas dificuldades, podem ser citadas:

1. A variedade de formato dos documentos, que mudam conforme o local e a época, complicando o *Spotting*.
2. A estrutura peculiar de discurso utilizada nas reivindicações: declarativa e de sentenças longas, que complicam a análise sintática.
3. A presença expressiva de termos de domínio específico de conhecimento, muitas vezes inéditos, que induzem a necessidade do uso de ontologias de patentes para identificação de tópicos e objetos secundários.
4. O aspecto referencial do discurso e o diálogo com elementos não textuais, como símbolos e ilustrações.

Este trabalho teve como objetivo possibilitar a captura de elementos textuais das reivindicações de patentes que explicitassem os termos, características e referências de interesse do ponto de vista de um examinador. Visando contornar as dificuldades (2), (3) e (4) acima apresentadas, tal captura foi feita através da análise de Funções Conceituais, na forma de segmentos semânticos. Os segmentos extraídos com o método desenvolvido tendem a conter informação relevante para a busca e sumarização de patentes.

Como principais contribuições do trabalho podem ser citadas:

- Um conjunto de atributos linguísticos que caracterizam de forma relativamente precisa as Funções Conceituais das reivindicações de patente;

- Um modelo de segmentação semântica extensível, que possibilita a criação e modificação de Funções Conceituais através de anotações no texto em linguagem natural;
- Um método supervisionado de extração de Funções Conceituais de reivindicações de patentes, na forma de segmentos semânticos;
- Um conjunto extensível de regras para extração de relações entre segmentos semânticos, na forma de grafos de relacionamento semântico.

Os resultados obtidos pelo sistema que implementa as contribuições citadas indica que o método é adequado para a extração de tópicos e referências, sendo indicado para a busca e sumarização de documentos. O desempenho obtido para Funções Conceituais de maior ambiguidade e a abrangência geral para o conjunto de reivindicações testado indica também que ainda há muito espaço para melhorias.

Como contribuição secundária, a arquitetura de processamento em linha de montagem (*pipeline*) desenvolvida para o método de extração expõe as falhas de cada etapa da decomposição do texto. Isto facilita a compreensão das deficiências de uma determinada etapa (e.g., POS-tagging) para o tipo de texto em análise, favorecendo a melhoria dos algoritmos utilizados. Como exemplo, citamos o caso do *feedback* do mWANN-Tagger [31] que permitiu melhorar o desempenho do analisador morfológico.

5.2 Trabalhos futuros

As principais melhorias a serem feitas estão relacionadas ao desempenho do classificador e ao algoritmo de hipótese e validação para segmentos. Uma das alternativas de classificação trata da divisão da WiSARD em um conjunto de redes menores, cada uma focada em um atributo diferente. O motivo da divisão é diminuir a influência que atributos sujeitos a maior quantidade de ruído têm no desempenho do classificador, quando todos os atributos são concatenados e mapeados na rede. Entretanto, esta alternativa implica na necessidade de resolver o problema de composição dos resultados das diferentes redes. O algoritmo de hipótese e validação precisa de filtros efetivos para evitar ao máximo a formulação de hipóteses erradas, algo que também sofre influência do classificador. A elaboração de tais filtros, possivelmente, passa pela criação de heurísticas baseadas nos conjuntos de pares hierárquicos e sintático-semânticos capturados na fase de treinamento do sistema. Para melhorar o tempo de resposta do sistema, além das melhorias já propostas, o código deve ser otimizado para eliminar operações redundantes ou desnecessárias. A aplicação de melhores técnicas computacionais, como paralelização e *caching* de resultados intermediários de certas operações também devem ser considerados.

Como forma de expandir o escopo do trabalho, técnicas de seleção de conteúdo e *spotting* podem ser incorporadas, tornando o processo de extração de informação totalmente automático. Além disto, a ausência de componentes específicos para análise de patentes permite que o sistema seja aplicado em tarefas de extração fora deste domínio. Tal não foi ainda experimentado. Outro ponto importante de expansão é a adoção do conhecimento de especialistas em patentes para a anotação das reivindicações, produzindo exemplos melhor alinhados com as necessidades reais de análise dos documentos de patente.

Referências Bibliográficas

[1] GHOULA, N., KHELIF, K., DIENG-KUNTZ, R. "Supporting patent mining by using ontology-based semantic annotations". In: *Web intelligence, IEEE/WIC/ACM international conference*, pp. 435–438, 2007.

[2] TADURI, S., LAU, G. T., LAW, K. H., et al. "A patent system ontology for facilitating retrieval of patent related information". In: *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance*, pp. 146–157, 2012.

[3] YANG, D. Y., SOO, V. M. "Extract conceptual graphs from plain texts in patent claims", *Engineering Applications of Artificial Intelligence*, v. 25, n. 4, pp. 874–887, 2012.

[4] BACH, N. X., MINH, N. L., OANH, T. T., et al. "A Two-Phase Framework for Learning Logical Structures of Paragraphs in Legal Articles", *ACM Transactions on Asian Language Information Processing (TALIP)*, v. 12, n. 1, pp. 3, 2013.

[5] FERREIRA, V. H., LOPES, L., VIEIRA, R., et al. "Automatic Extraction of Domain Specific Non-taxonomic Relations from Portuguese Corpora". In: *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE/WIC/ACM International Joint Conferences*, v. 3, pp. 135–138, 2013.

[6] BRUCKSCHEN, M., DE SOUZA, J. G. C., VIEIRA, R., et al. "Sistema SeRELeP para o reconhecimento de relaç oes entre entidades mencionadas", *Mota and Santos (Mota and Santos, 2008)*, 2008.

[7] CAPUTO, G. M. *Sistema Computacional para o processamento textual de patentes industriais*. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil, 2006.

[8] GIRJU, R., BADULESCU, A., MOLDOVAN, D. "Automatic discovery of part-whole relations", *Computational Linguistics*, v. 31, n. 1, pp. 83–135, 2006.

[9] PANTEL, P., PENNACCHIOTTI, M. "Espresso: Leveraging generic patterns for automatically harvesting semantic relations". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 113–120. Association for Computational Linguistics, 2006.

[10] MINH, N. L., SHIMAZU, A. "A semi supervised learning model for mapping sentences to logical forms with ambiguous supervision", *Data & Knowledge Engineering*, v. 90, pp. 1–12, 2014.

[11] GRUBER, T. "Ontology". In: Liu, L., Özsu, M. T. (Eds.), *Encyclopedia of Database Systems*, 1 ed., Springer-Verlag, 2008.

[12] ROSSE, C., MEJINO, J. L. V. "The Foundational Model of Anatomy Ontolog". In: A. Burger, D. D., Baldock, R. (Eds.), *Anatomy Ontologies for Bioinformatics: Principles and Practice*, 1 ed., pp. 59–117, London, Springer, 2007.

[13] SCHRIML, L. M., ARZE, C., NADENDLA, S., et al. "Disease Ontology: a backbone for disease semantic integration", *Nucleic acids research*, v. 40, jan. 2012.

[14] FAO. *FAO ontology http://aims.fao.org/geopolitical.owl*. Relatório técnico, Food and Agriculture Organization of the United Nations (FAO).

[15] BRICKLEY, D., MILLER, L. *Friend Of A Friend (FOAF) specification http://xmlns.com/foaf/spec/*. Relatório técnico, FOAF Project – http://www.foaf-project.org/.

[16] PEASE, A., NILES, I., LI, J. "The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications". In: *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada*, Edmonton, Canadá, jul. 2002.

[17] GRENON, P., SMITH, B. "SNAP and SPAN: Towards Dynamic Spatial Ontology", *Spatial Cognition and Computation*, v. 4, n. 1, pp. 69–103, 2004.

[18] W3C. *OWL Web Ontology Language: http://www.w3.org/TR/owl-ref/*. Relatório técnico, World Wide Web Consortium (W3C), .

[19] W3C. *Extensible Markup Language: http://www.w3.org/TR/rec-xml*. Relatório técnico, World Wide Web Consortium (W3C), .

[20] KRÖTZSCH, M., SIMANCIK, F., HORROCKS, I. "Description Logics." *IEEE Intelligent Systems*, v. 29, n. 1, pp. 12–19, 2014.

[21] W3C. *Resource Description Framework: http://www.w3.org/TR/PR-rdf-syntax/*. Relatório técnico, World Wide Web Consortium (W3C), .

[22] BIZER, C., LEHMANN, J., KOBILAROV, G., et al. "DBpedia - A Crystallization Point for the Web of Data", *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 7, n. 3, pp. 154–165, 2009.

[23] MILLER, G. A., BECKWITH, R., FELLBAUM, C. D., et al. "WordNet: An online lexical database", *Int. J. Lexicograph*, v. 3, n. 4, pp. 235–244, 1990.

[24] PETROV, S., DAS, D., MCDONALD, R. "A Universal Part-of-Speech Tagset". In: *Proceedings of 8th International Conference on Language Resourses and Evaluation (LREC)*, maio 2012.

[25] NASEEM, T., SNYDER, B., EISENSTEIN, J., et al. "Multilingual Part-Of-Speech Tagging: Two Unsupervised Approaches", *Journal of Artificial Intelligence Research*, v. 36, pp. 341–385, 2009.

[26] DEROSE, S. J. "Grammatical category disambiguation by statistical optimization", *Computational Linguistics*, v. 14, n. 1, pp. 31–39, 1988.

[27] CHARNIAK, E. "Statistical Techniques for Natural Language Parsing", *AI Magazine*, v. 18, n. 4, pp. 33–44, 1997.

[28] MERIALDO, B. "Tagging English text with a probabilistic model", *Computational Linguistics*, v. 20, n. 2, pp. 155–171, 1994.

[29] LOFTSSON, H. "Tagging a morphologically complex language using heuristics", *Advances in Natural Language Processing*, pp. 640–651, 2006.

[30] GIMENEZ, J., MARQUEZ, L. "A general POS tagger generator based on Suport Vector Machines". In: *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, pp. 43–46, 2004.

[31] CARNEIRO, H. C. C., FRANÇA, F. M. G., LIMA, P. M. V. "WANN-Tagger: A Weightless Artificial Neural Network Tagger for the Portuguese Language". In: *Proceedings of the Intenational Conference on Fuzzy Computation and International Conference on Neural Computation*, pp. 330–335, out 2010.

[32] MANNING, C. D. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In: Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, v. 6608, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 171–189, 2011. doi: 10.1007/978-3-642-19400-9_14. Disponível em: <http://dx.doi.org/10.1007/978-3-642-19400-9_14>.

[33] AZEREDO, J. C. *Iniciação a Sintaxe do Portugues*. Zahar, 2001.

[34] CHOMSKY, N. *Syntactic Structures*. Mouton, 1957.

[35] MANNING, C. D., SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[36] GILDEA, D., JURAFSKY, D. "Automatic Labeling of Semantic Roles". In: *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00)*, pp. 512–520, Hong Kong, out 2000.

[37] ALAN, K. *Natural Language Semantics*. Oxford, Blackwell Publishers Ltd, 2001.

[38] LYONS, J. *Introduction to Theoretical Linguistics*. Cambridge University Press, 1968.

[39] CRUSE, A. D. *Lexical semantics*. Cambridge University Press, 1986.

[40] BEAN, A., GREEN, R. *Relationships in the Organization of Knowledge*. Fundamental Theories of Physics. Springer, 2001.

[41] KAMP, H. "A theory of truth and semantic representation". In: J.A.G. Groenendijk, T. J., Stokhof, M. (Eds.), *Formal Methods in the Study of Language*, 1 ed., pp. 277–322, Amsterdam, Mathematical Centre Tracts 135, 1981.

[42] JACKENDOFF, R. *Semantic Structures*. Current Studies in Linguistics. MIT Press, 1992.

[43] KLAS, W., SCHREFL, M. "Semantic data modeling". In: *Metaclasses and Their Application*, v. 943, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 1995.

[44] KATZ, J. J., FODOR, J. A. "The structure of a semantic theory", *Language*, pp. 170–210, 1963.

[45] MOLDOVAN, D., BADULESCU, A., TATU, M., et al. "Models for the semantic classification of noun phrases". In: *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pp. 60–67, 2004.

[46] HURTADO, L., SEGARRA, E., GARCÍA, F., et al. "Language Understanding Using n-multigram Models". In: Vicedo, J., Martínez-Barco, P., Muńoz, R., et al. (Eds.), *Advances in Natural Language Processing*, v. 3230, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 207–219, 2004.

[47] CHINCHOR, N., ROBINSON, P. "MUC-7 named entity task definition". In: *Proceedings of the 7th Conference on Message Understanding*, 1997.

[48] FINKEL, J. R., GRENAGER, T., MANNING, C. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". In: *43rd Annual Meeting of the Association for Computational Linguistics*, p. 363–370, 2005.

[49] NOTHMAN, J., RINGLAND, N., RADFORD, W., et al. "Learning multilingual named entity recognition from Wikipedia", *Artificial Intelligence 194*, p. 151–175, 2013.

[50] PEREIRA, B. O. *Resolução de Entidades Nomeadas utilizando recursos em Linked Data*. Dissertação de mestrado, PPGI - Programa de pós-graduação em Informática - IM - Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2012.

[51] AMORIM, C., SOUSA, C. *Gramática da Língua Portuguesa: 3º Ciclo do Ensino Básico e Ensino Secundário*. Areal Editores, 2012.

[52] CARDIE, C., WAGSTAÆ, K. "Noun phrase coreference as clustering". In: *Proceedings of the Joint Sigdat Conference on empirical methods in natural language processing and very large corpora*, p. 82–89, New Brunswick, NJ, EUA, 1999.

[53] CUEVAS, R. R. M., HONDA, W. Y., LUCENA, D. J., et al. "Portuguese Pronoun Resolution: Resources and Evaluation". In: Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*, v. 4919, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 344–350, 2008.

[54] CHAVES, A. R., RINO, L. H. M. "A resolução de pronomes anafóricos do português com base em heurísticas que apontam o antecedente". In: *Congresso*

de Pós-Graduação, Universidade Federal de São Carlos, São Carlos, SP, Brasil, 2007.

[55] FUNDEL, K., KÜFFNER, R., ZIMMER, R. "RelEx—Relation extraction using dependency parse trees", *Bioinformatics*, v. 23, n. 3, pp. 365–371, 2007.

[56] CHUN, H. W., TSURUOKA, Y., KIM, J. D., et al. "Extraction of gene-disease relations from Medline using domain dictionaries and machine learning." In: *Pacific Symposium on Biocomputing*, v. 11, pp. 4–15, 2006.

[57] AUGER, A., BARRIÈRE, C. "Pattern-based approaches to semantic relation extraction: A state-of-the-art", *Terminology*, v. 14, n. 1, pp. 1–19, 2008.

[58] MCCULLOCH, W. S., PITTS, W. "A logical calculus of the ideas immanent in nervous activity", *The bulletin of mathematical biophysics*, v. 5, n. 4, pp. 115–133, 1943.

[59] ROSENBLATT, F. *Principles of neurodynamics*. Spartan Book, 1962.

[60] ALEKSANDER, I., MORTON, H. *An introduction to neural computing*, v. 240. Chapman and Hall London, 1990.

[61] WASSERMAN, P. D. *Neural computing: theory and practice*. Van Nostrand Reinhold Co., 1989.

[62] MINSKY, M., PAPERT, S. "Perceptrons", *Cambridge, Ma*, 1969.

[63] BLEDSOE, W. W., BROWNING, I. *Pattern recognition and reading by machine*. PGEC, 1959.

[64] ALEKSANDER, I. "Self-adaptive universal logic circuits", *Electronics Letters*, v. 2, n. 8, pp. 321–322, 1966.

[65] ALEKSANDER, I., THOMAS, W. V., BOWDEN, P. A. "WISARD· a radical step forward in image recognition", *Sensor review*, v. 4, n. 3, pp. 120–124, 1984.

[66] KANERVA, P. *Sparse distributed memory*. MIT press, 1988.

[67] FAIRHURST, M. C., BISSET, D. L., OTHERS. "Adaptive pattern recognition using goal seeking neurons", *Pattern recognition letters*, v. 12, n. 3, pp. 131–138, 1991.

[68] ALEKSANDER, I. "Ideal neurons for neural computers", *Parallel Processing in Neural Systems and Computers*, pp. 225–228, 1990.

[69] MRSIC-FLOGEL, J. "Convergence Properties of Self-Organizing Maps". In: *Proceedings of the International Conference on Artificial Neural Networks*, pp. 879–886, Amsterdan, 1991.

[70] CARDOSO, D. O., CARVALHO, D. S., ALVES, D. S. F., et al. "Credit analysis with a clustering RAM-based neural classifier". In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 517–522, Bruges, Bélgica, abr. 2014.

[71] SOARES, C. M., SILVA, C. L. F., DE GREGORIO, M., et al. "Uma Implementação em Software do Classificador WISARD". In: *Anais do V Simpósio Brasileiro de Redes Neurais*, pp. 225–229, Belo Horizonte, Brasil, 1998.

[72] COUTINHO, P. V. S., CARNEIRO, H. C. C., CARVALHO, D. S., et al. "Extracting rules from DRASiW s 'mental images"'. In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Bélgica, abr. 2014.

[73] GRIECO, B., LIMA, P. M. V., DE GREGORIO, M., et al. "Producing pattern examples from "mental" images", *Neurocomputing*, v. 73, n. 7, pp. 1057–1064, 2010.

[74] CARVALHO, D. S., CARNEIRO, H. C. C., FRANÇA, F. M. G., et al. "B-bleaching: Agile Overtraining Avoidance in the WiSARD Weightless Neural Classifier". In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Bélgica, abr. 2013.

[75] SOUZA, C. R., NOBRE, F. F., LIMA, P. M. V., et al. "Recognition of HIV-1 subtypes and antiretroviral drug resistance using weightless neural networks". In: *Proc. of ESANN 2012*, pp. 429–434. i6doc.com, abr. 2012.

[76] KAVZOGLU, T., MATHER, P. M. "Pruning artificial neural networks: an example using land cover classification of multi-sensor images", *International Journal of Remote Sensing*, v. 20, n. 14, pp. 2787–2803, 1999.

[77] ESPOSITO, F., MALERBA, D., SEMERARO, G., et al. "A comparative analysis of methods for pruning decision trees", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 19, n. 5, pp. 476–491, 1997.

[78] SILVA, J., BRANCO, A., CASTRO, S., et al. "Out-of-the-Box Robust Parsing of Portuguese". In: *Proceedings of the 9th International Conference on the Computational Processing of Portuguese PROPOR'10*, pp. 75–85, 2010.

[79] BIRD, S., KLEIN, E., LOPER, E. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[80] SMITH, R. "An Overview of the Tesseract OCR Engine", *ICDAR*, v. 7, pp. 629–633, 2007.

[81] MARCUS, M., MARCINKIEWICZ, M. A., SANTORINI, B. "Building a large annotated corpus of English: The Penn Treebank", *Computational linguistics*, v. 19, n. 2, pp. 313–330, 1993.

[82] FREITAS, A., CARVALHO, D. S., DA SILVA, J. C. P., et al. "A Semantic Best-Effort Approach for Extracting Structured Discourse Graphs from Wikipedia". In: *Proc. of the 1st Workshop on the Web of Linked Entities,(ISWC)*, 2012.

[83] CARVALHO, D. S., FREITAS, A., DA SILVA, J. C. P. "Graphia: Extracting Contextual Relation Graphs from Text". In: *The Semantic Web: ESWC 2013 Satellite Events*, Springer, pp. 236–241, 2013.

[84] VAN LINT, J. H. "A theorem on equidistant codes", *Discrete Mathematics*, v. 6, n. 4, pp. 353–358, 1973.

[85] BOGDANOVA, G. T., ZINOVIEV, V. A., TODOROV, T. J. "On the construction of q-ary equidistant codes", *Problems of Information Transmission*, v. 43, n. 4, pp. 280–302, 2007.

[86] MINDER, L., SAUERWALD, T., WEGNER, S. A. "Asymptotic bounds on the equilateral dimension of hypercubes", *Graphs and Combinatorics*, pp. 1–8, 2014.

[87] DE GREGORIO, M. "Is that Portal Gothic? A Hybrid System for Recognising Architectural Portal Shapes", 1996.

[88] SANTOS, C. M. . D. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[89] DAIBER, J., JAKOB, M., HOKAMP, C., et al. "Improving Efficiency and Accuracy in Multilingual Entity Extraction". In: *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[90] CARVALHO, D. S., ÇALLI, C., FREITAS, A., et al. "EasyESA: A Low-effort Infrastructure for Explicit Semantic Analysis". In: *Proceedings of the 13th International Semantic Web Conference (ISWC)*, 2014.

[91] DE PAIVA, V., RADEMAKER, A., DE MELO, G. "OpenWordNet-PT: An Open Brazilian WordNet for Reasoning". In: *Proceedings of the 24th International Conference on Computational Linguistics*, 2012. Disponível em: <http://hdl.handle.net/10438/10274>.

[92] QUINLAN, J. R. *C4.5: programs for machine learning*, v. 1. Morgan kaufmann, 1993.

[93] COHEN, W. W. "Fast Effective Rule Induction". In: *Twelfth International Conference on Machine Learning*, pp. 115–123. Morgan Kaufmann, 1995.

[94] CORTES, C., VAPNIK, V. "Support-vector networks", *Machine learning*, v. 20, n. 3, pp. 273–297, 1995.

[95] HAND, D. J., YU, K. "Idiot's Bayes—not so stupid after all?" *International statistical review*, v. 69, n. 3, pp. 385–398, 2001.

[96] HALL, M., FRANK, E., HOLMES, G., et al. "The WEKA data mining software: an update", *ACM SIGKDD explorations newsletter*, v. 11, n. 1, pp. 10–18, 2009.

[97] CHANG, C. C., LIN, C. J. "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, v. 2, pp. 27:1–27:27, 2011. Software disponível em http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Apêndice A

Tabelas complementares

Tabela A.1: Mapeamento de classes gramaticais (POS-tags) do mWANN-Tagger para o LX-Tagger

Tag mWANN	Tag LX
V (Verbo)	V
N (Substantivo [Nome])	N
ADP (Adposição)	P (Preposição)
ADJ (Adjetivo)	A
DET (Determinante)	ART (Artigo)
PUNC (Pontuação)	PNT
PRON (Pronome)	PRS (Pronome não reflexivo)
NUM (Número)	CARD (Cardinal)
ADV (Advérbio)	ADV
CJ (Conjunção)	CONJ

Tabela A.2: Conjunto de regras para extração de relacionamentos semânticos

Sequência	Sujeito	Objeto	Relação (predicado)
ASSUNTO_PAT, REF, REF_REIVIND	ASSUNTO_PAT	REF_REIVIND	de acordo com
REF_REIVIND, REF_REIVIND_NUM	REF_REIVIND_NUM	REF_REIVIND	identifica
ASSUNTO_PAT, *, CARACT_ASSUNTO, *, OBJ_PAT	ASSUNTO_PAT	OBJ_PAT	(verbo usado em CARACT_ASSUNTO)
OBJ_PAT, CARACT_OBJ, OBJ_PAT	OBJ_PAT	OBJ_PAT	(verbo usado em CARACT_OBJ)
OBJ_PAT, ILUST_REF	ILUST_REF	OBJ_PAT	ilustra
OBJ_PAT, CARACT_OBJ, [^OBJ_PAT][1]	CARACT_OBJ	OBJ_PAT	caracteriza

Apêndice B

Configurações da ferramenta WEKA utilizadas nos experimentos

Perceptron Multicamada

GUI	False
autoBuild	True
debug	False
decay	False
hiddenLayers	a
learningRate	0.3
momentum	0.2
nominalToBinaryFilter	True
normalizeAttributes	True
normalizeNumericClass	True
reset	True
seed	0
trainingTime	500
validationSetSize	0
validationThreshold	20

SVM

SVMType	C-SVC (classification)
cacheSize	40.0
coef0	0.0
cost	1.0
debug	False
degree	3
doNotReplaceMissingValues	False
eps	0.001
gamma	0.0
kernelType	linear: u'*v
loss	0.1
normalize	False
nu	0.5
probabilityEstimates	False
seed	1
shrinking	True
weights	

C4.5

binarySplits	False
confidenceFactor	0.25
debug	False
minNumObj	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False

RIPPER

checkErrorRate	True
debug	False
folds	3
minNo	2.0
optimizations	2
seed	1
usePruning	True

Naive Bayes

debug	False
displayModelInOldFormat	False
useKernelEstimator	False
useSupervisedDiscretization	False

Apêndice C

Patente referência para os exemplos: Blindagem Protetora contra Arrombamento de Cofres

República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) **PI0803602-0 A2**

* B R P I 0 8 0 3 6 0 2 A 2 *

(22) Data de Depósito: 27/06/2008
(43) Data da Publicação: 31/05/2011
(RPI 2108)

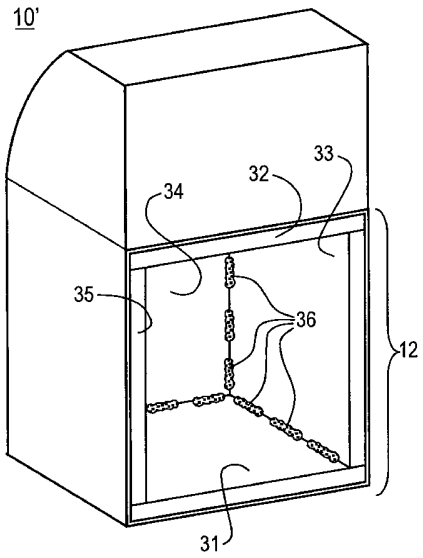
(51) Int.Cl.:
E05G 1/024 2006.01
E05G 1/10 2006.01

(54) Título: **BLINDAGEM PROTETORA CONTRA ARROMBAMENTO DE COFRES**

(73) Titular(es): ITAUTEC S.A.- GRUPO ITAUTEC

(72) Inventor(es): Ronaldo Marques, Vanderley de Assis Reis

(57) Resumo: BLINDAGEM PROTETORA CONTRA ARROMBAMENTO DE COFRES Blindagem protetora contra arrombamento de cofres compreendendo um conjunto de painéis de blindagem (31, 32, 33, 34, 35), justapostos às faces internas das paredes, piso e teto do cofre e unidos solidamente entre si ao Longo de suas bordas mediante soldagem (36). Cada painel está formado por várias camadas protetoras, compreendendo proteção química (38), proteção mecânica e proteção contra perfuração (39) contra ferramentas cortantes. A proteção química compreende substâncias que, sob a ação da chama de um maçarico, desprendem compostos voláteis agressivos às mucosas do operador, enquanto a proteção contra perfliração compreende uma camada de material de elevada dureza (39a), superior a 500 1-1V 1. Opcionalmente, cada painel pode estar provido de uma placa de aço inoxidável (42).



PI0803602-0

"BLINDAGEM PROTETORA CONTRA ARROMBAMENTO DE COFRES"

Campo da invenção

Refere-se a presente invenção à proteção de equipamentos acondicionadores de valores e, mais particularmente, à proteção dos cofres utilizados em terminais de atendimento bancário – ATM's.

Descrição do estado da técnica

Segundo é do conhecimento geral, as instituições bancárias disponibilizam em suas agências ou em locais de grande movimento – tais como supermercados, postos de abastecimento, aeroportos, etc. – equipamentos do tipo ATM que permitem executar diversos tipos de operações financeiras, tais como pagamentos, consulta de extratos e, principalmente, saques em numerário através do uso de cartões e/ou senhas. Evidentemente, para atender a esta última função, os ATM's necessitam armazenar um considerável estoque de cédulas, as quais são acondicionadas em magazines próprios que, por sua vez, estão colocados em uma parte específica do gabinete destes ATM's comumente chamadas de cofres.

A Fig. 1 ilustra um modelo típico de ATM 10, no qual a porção superior 11do gabinete compreende os dispositivos de interface com o usuário, tais como leitor de cartões, teclado, tela de vídeo, etc. O cofre que armazena os valores ocupa a porção inferior 12 do gabinete, e sua comunicação com os dispositivos de interface é feita através de uma fenda relativamente estreita, cujas dimensões e posição não permitem a extração das cédulas ali armazenadas. Como o arrombamento destes cofres constitui o método mais comum utilizado para a extração dos valores, tais cofres são fabricados com paredes reforçadas de modo a dificultar a ação dos assaltantes.

O documento de patente PI0104638-1 intitulado "Sistema de segurança e blindagem para equipamentos acondicionadores de valores e equivalentes" descreve uma armadura de segurança aplicada externamente ao

cofre, composta por paredes frontal 13, traseira 14 e laterais 15, 16 conforme ilustrado na Fig. 2. De acordo com o que mostra a Fig. 3, cada uma destas paredes está formada por duas chapas de aço carbono de 19mm de espessura 17, 18, paralelamente dispostas, sendo o espaço entre estas preenchido com argamassa de cimento aditivado com óxido de alumínio 19 para aumentar sua dureza. As faces externas destas paredes estão revestidas com uma blindagem química 21 composta por uma tinta impregnada com piche e enxofre. Ao ser submetida a um calor intenso, tal como o proveniente de um maçarico, esta composição desprende fumaça e gases tóxicos tornando impossível a permanência, no local, do eventual arrombador.

Apesar de oferecer alguma proteção contra a ação dos meliantes, o objeto deste pedido apresenta alguns inconvenientes, o primeiro dos quais é o fato da blindagem química 21 estar situada sobre a superfície externa das paredes protetoras. Isto facilita a sua remoção por meio de solvente e espátula, neutralizando assim a sua finalidade.

O pedido de patente PI0403799 intitulado "Metodologia para reforço de instalações e equipamentos bancários" prevê a instalação de jaquetas externas nos ATM's e similares, compreendendo painéis ou jaquetas de proteção em forma de caixas achatadas feitas em chapa metálica, que são montadas externamente junto às laterais do equipamento. Conforme ilustrado na Fig. 4, estas painéis apresentam internamente uma primeira camada de material que propicia resistência ao ataque com maçarico, essencialmente constituída por um composto inflamável contendo betume, breu e enxofre. A proteção mecânica é propiciada por uma camada 24 de concreto armado com barras de aço 25, contendo, além dos componentes usuais na confecção do concreto, quantidades especificadas de Koridon e Dramix, o primeiro sendo um abrasivo à base de trióxido de alumínio e o segundo constituído por numerosas fibras de aço. A fixação destas jaquetas ao gabinete é feita através de parafusos e/ou porcas, pela parte interna do cofre.

Os sistemas de blindagem apresentados nos dois documentos citados apresentam a desvantagem de serem facilmente detectados pelos arrombadores, permitindo que estes se equipem antecipadamente com meios, ferramentas e dispositivos capazes de penetrar as referidas painéis de blindagem.

Objetivos da invenção

Em vista do exposto, constitui o objetivo principal da presente invenção o provimento de um sistema de blindagem para cofres capaz de prover uma adequada proteção contra tentativas de arrombamento.

Constitui outro objetivo o provimento de um sistema de blindagem cuja existência não seja imediatamente percebida pelos eventuais arrombadores.

Descrição resumida da invenção

Os objetivos acima, bem como outros, são atingidos pela invenção mediante o provimento de um sistema que compreende um conjunto de painéis de blindagem instaladas internamente ao cofre, mais especificamente, justapostas às faces internas das paredes, piso e teto do cofre e unidas solidamente entre si ao longo de suas bordas.

De acordo com outra característica da invenção, os ditos painéis de blindagem são unidos entre si por meios de travamento mútuo totalmente inacessíveis externamente.

De acordo com outra característica da invenção, os ditos meios de travamento mútuo compreendem cordões de solda ao longo dos diedros reentrantes formados pelos ditos painéis em contato ao longo de suas bordas.

De acordo com outra característica da invenção, o conjunto de painéis de blindagem forma, após a união de seus componentes, uma caixa blindada inserida no interior do cofre.

De acordo com outra característica da invenção, cada painel de blindagem compreende pelo menos uma camada de proteção mecânica, pelo

menos uma camada de proteção química e pelo menos uma camada de proteção contra perfuração por ferramentas cortantes.

De acordo com outra característica da invenção, a proteção química compreende materiais que, ao serem aquecidos mediante aplicação de chama de maçarico, desprendem gases ou vapores agressivos, irritantes ou tóxicos.

De acordo com outra característica da invenção, a camada de proteção contra perfuração compreende materiais abrasivos de elevada dureza.

De acordo com outra característica da invenção, os ditos materiais de elevada dureza compreendem carbonetos complexos de cromo, nióbio ou boro.

Descrição das figuras

As demais características e vantagens da invenção tornar-se-ão mais evidentes a partir da descrição de uma concretização exemplificativa e não limitativa, e das figuras que a ela se referem, nas quais:

A figura 1 ilustra o aspecto externo de um terminal de auto-atendimento bancário (ATM) convencional.

A figura 2 ilustra a técnica conhecida descrita no documento de patente PI0104638-1, consistindo na aplicação de uma armadura externamente às paredes do cofre.

A figura 3 mostra em detalhe a estrutura dos painéis utilizados na armadura da figura anterior.

A figura 4 ilustra outra técnica conhecida, descrita no documento de patente PI0403799.

A figura 5 ilustra um cofre provido de painéis de blindagem interna de acordo com os princípios da invenção.

A figura 6a mostra a constituição do painel de blindagem, de acordo com os princípios da invenção.

A figura 6b mostra uma forma alternativa do painel de blindagem, de acordo com os princípios da invenção.

A figura 6c mostra o comportamento do painel produzido de acordo com os princípios da invenção, ao ser submetido à tentativa de perfuração por meio de um maçarico.

Descrição detalhada da invenção

Considerando que não existe blindagem absolutamente inviolável, o objetivo da colocação de blindagens em cofres e assemelhados tem como propósito principal dificultar o seu arrombamento, mediante o provimento de estruturas de proteção que retardem, tanto quanto possível, o acesso ao seu interior. Com efeito, os locais em que se encontram os cofres que armazenam valores estão equipados com equipamentos destinados à detecção de anormalidades, entre as quais se contam as tentativas de arrombamento ou violação. Conseqüentemente, o início das atividades de arrombamento será imediatamente detectado e informado à central de supervisão, permitindo acionar as equipes de segurança para as devidas providências. Uma vez que o deslocamento destas equipes demanda um certo período de tempo, que pode chegar a uma hora ou até mais, a função básica das blindagens consiste em fazer com que o tempo necessário à realização do arrombamento se torne suficientemente longo, o que permitirá a chegada das equipes de segurança antes que seja consumado o furto.

O sistema de blindagem proposto pela presente invenção atinge plenamente este objetivo, através dos seguintes recursos:

- A instalação da blindagem no interior do cofre não altera seu aspecto externo, que permanece idêntico àquele do cofre não blindado. Em consequência, os meliantes são levados a pensar que o cofre poderia ser arrombado por meio de ferramentas e utensílios convencionais, e somente após iniciar a tentativa de arrombamento perceberão que a natureza da proteção exige equipamentos de maior poder de penetração, os quais

geralmente não estão ao alcance imediato dos criminosos, levando-os a abandonar a tentativa de furto.

- Ademais, quando comparada com os painéis pertencentes ao estado da técnica, a blindagem proposta é mais eficaz, ou seja, mesmo com o uso de equipamentos de arrombamento mais sofisticados, sua violação é difícil e muito demorada.

Fazendo referência, agora, à Fig. 5, o cofre 12' do ATM 10' está blindado de acordo com a invenção, mediante o provimento dos painéis de blindagem de fundo 31, de teto 32, laterais 33 e 35 e frontal 34 justapostos às faces internas do gabinete do cofre. Para maior clareza de representação, não foi ilustrada a porta do cofre, a qual também possui um ou mais painéis de blindagem cuja estrutura é similar àquela dos painéis citados.

A instalação da blindagem compreende o posicionamento dos painéis conforme indicado na figura, seguindo-se a união mútua destes de forma sólida, por meio dos cordões de solda 36, o conjunto formando um receptáculo ou caixa rígida, em cujo interior são acondicionados os magazines contendo os valores, bem como demais mecanismos que provém a movimentação das cédulas entre a porção superior do ATM e o cofre.

A Fig. 6a ilustra uma primeira concretização preferida da estrutura dos referidos painéis, que compreende, no sentido de fora para dentro, uma chapa de acondicionamento externa 37 em aço, com espessura entre 1,2mm e 3mm, pelo menos uma camada de proteção química 38 com espessura entre 5mm e 20mm, seguindo-se pelo menos uma placa de alta dureza 39 com espessura entre 8mm e 15mm e a chapa de acondicionamento interna 41 de aço, similar à dita chapa externa, com espessura entre 1,2mm e 3mm.

A camada de proteção química 38 é inflamável, e compreende uma matriz de natureza asfáltica que se encontra preenchida com grânulos de materiais que, ao queimar, libertam substâncias agressivas as quais, em

contato com as mucosas do operador, produzem intenso desconforto, obrigando-o a desistir de seu intento.

A placa de alta dureza 39 constitui a proteção contra perfuração por meio de brocas ou assemelhados, e compreende uma base metálica 39b, preferencialmente uma chapa de aço, revestida com uma camada 39a de liga de alta dureza contendo inclusões de carbonetos e boretos extra duros de cromo, nióbio e boro, numa proporção igual ou superior a 50% destes compostos. A dureza de dita camada deve ser superior a 500 HV1, aproximadamente equivalente a 50 HRc. A espessura da chapa de base 39b varia entre 5mm e 10mm, e a da camada de revestimento entre 3mm e 5mm. Numa concretização preferida da invenção, utiliza-se as placas CastoDur produzidas pela Eutectic Castolin, cuja camada de revestimento possui dureza entre 60 e 65 HRc. Outros materiais equivalentes, de outros fornecedores, poderão ser utilizados, desde que apresentando características de dureza e resistência à abrasão iguais ou superiores àquela mencionada.

Segundo mostra a Fig. 6a, o conjunto que forma o painel é fechado internamente por uma chapa de aço 41, com espessura entre 1,2mm e 3mm, na qual podem ser fixados, por exemplo mediante soldagem, os elementos de suporte dos magazines e equipamentos auxiliares do terminal de atendimento, tais como trilhos, braçadeiras, etc.

Na Fig. 6b está ilustrada uma concretização alternativa do painel de blindagem, no qual se intercalou pelo menos uma chapa de aço inoxidável 42 entre a placa 39 e a chapa interna 41, dita chapa de aço inoxidável tendo uma espessura entre 2,5mm e 6mm. Preferencialmente, o dito aço inoxidável é do tipo AISI 304.

A Fig. 6c exemplifica o efeito protetor da camada 38 quando se utiliza a chama de um maçarico 43 para tentar perfurar uma armadura idêntica àquela ilustrada na Fig. 6a. Segundo ilustrado, o calor da chama resulta na produção de uma concavidade 44 que reflete a chama, que volta na

direção do operador, conforme indicado pelas setas 45. A referida concavidade resulta do efeito da fusão e da queima do material da camada 38. Segundo mencionado, um dos componentes dessa camada é uma matriz de natureza asfáltica, podendo-se utilizar uma variedade de materiais, tais como o betume, o piche e similares, os quais se fundem quando moderadamente aquecidos, o que possibilita sua aplicação mediante despejamento. As altas temperaturas do maçarico produzem a sua ignição, bem como a queima dos demais componentes incluídos nesta camada sob forma granulada. Um destes componentes é a resina vegetal, por exemplo o breu ou colofônio, que produz fumaça durante sua queima. O outro componente pode ser de origem mineral, tal como o enxofre que produz SO_2 (óxido sulfuroso) quando aquecido, o cloreto de amônio cujo aquecimento produz NH_3 (amônia) e HCl (ácido clorídrico), etc., podendo-se, ainda, utilizar partículas ou grânulos de materiais orgânicos que libertam compostos voláteis cáusticos ou agressivos sob a ação do calor. Em resumo, o intenso aquecimento devido à chama do maçarico produz a combustão dos materiais que formam a camada 38, o que intensifica a chama refletida em direção ao operador do maçarico. Esta combustão é acompanhada da produção de substâncias voláteis – gases, vapores e fumaças – 46 altamente prejudiciais às mucosas dos olhos, nariz, garganta e sistema respiratório, obrigando o arrombador a se afastar ou, mesmo, abandonar o local. Evidentemente, um efeito similar ocorre ao se aplicar a chama do maçarico à armadura exemplificada na Fig. 6b.

Segundo mencionado e ilustrado, a instalação interna dos painéis apresenta a vantagem de permitir o *retrofitting*, ou seja, a sua instalação posterior em ATM's ou cofres já existentes, com um mínimo de dificuldade, dispensando, até, a perfuração das paredes dos gabinetes para montagem dos painéis.

Se bem que a invenção tenha sido descrita fazendo referência a uma concretização exemplificativa específica, fica entendido que

modificações e alterações poderão ser feitas por técnicos no assunto, permanecendo dentro do espírito e âmbito da idéia inventiva. Assim, por exemplo, apesar da blindagem estar ilustrada na Fig. 5 como sendo composta de painéis inteiriços, com as mesmas dimensões das paredes laterais, de frente, de topo e de fundo do cofre, como indicado, tais painéis poderão estar compostos pela união de sub-painéis de menores dimensões, sem ultrapassar os limites da invenção.

Outrossim, os painéis de blindagem das figuras 6a e 6b correspondem a concretizações preferidas, nas quais se tem, pela seguinte ordem, de fora para dentro:

- chapa de acondicionamento 37;
- camada de proteção química 38;
- placa de alta dureza 39;
- chapa opcional de aço inoxidável 42;
- chapa de acondicionamento 41.

Todavia, as chapas e camadas que, em conjunto, constituem os painéis de blindagem da invenção podem estar dispostas em uma ordem diferente daquelas ilustradas nas concretizações preferidas, desde que sejam mantidas em ambas as faces externas, as chapas de acondicionamento 37 e 41 e a camada de alta dureza 39a fique voltada para o lado de fora do cofre.

De acordo, a invenção se encontra definida e delimitada pelo conjunto de reivindicações que se segue.

REIVINDICAÇÕES

1. Blindagem protetora contra arrombamento de cofres compreendendo um conjunto de painéis de blindagem (31, 32, 33, 34, 35) caracterizada pelo fato de ditos painéis estarem instalados no interior do cofre, mediante justaposição às faces internas das paredes, piso e teto do cofre.

2. Blindagem protetora de acordo com a reivindicação 1, caracterizada pelo fato de ditos painéis estarem solidamente unidos entre si ao longo de suas bordas.

3. Blindagem protetora de acordo com a reivindicação 2, caracterizada pelo fato de os meios de união entre as bordas dos painéis adjacentes serem totalmente inacessíveis a partir do exterior do cofre.

4. Blindagem protetora de acordo com a reivindicação 3, caracterizada pelo fato de ditos meios de união compreenderem cordões de solda (36) ao longo dos diedros reentrantes formados ao longo das bordas das painéis em contato mútuo.

5. Blindagem protetora de acordo com qualquer uma das reivindicações anteriores, caracterizada pelo fato de cada painel de blindagem compreender:

- uma chapa de acondicionamento externa (37);
- pelo menos uma camada de proteção química (38);
- pelo menos uma placa de alta dureza (39);
- uma chapa de acondicionamento interna (41)

6. Blindagem protetora de acordo com a reivindicação 5, caracterizada pelo fato de dita pelo menos uma camada de proteção química (38) compreender materiais que libertam compostos voláteis, cáusticos ou agressivos (46), sob a ação do calor da chama do maçarico (43).

7. Blindagem protetora de acordo com as reivindicações 5 ou 6, caracterizada pelo fato de dita pelo menos uma camada de proteção química (38) compreender uma matriz de natureza asfáltica contendo

grânulos de materiais resinosos.

8. Blindagem protetora de acordo com as reivindicações 5 ou 6, caracterizada pelo fato de dita pelo menos uma camada de proteção química compreender grânulos de substâncias inorgânicas.

9. Blindagem protetora de acordo com as reivindicações 6 ou 7, caracterizada pelo fato de dita matriz de natureza asfáltica compreender uma ou mais substâncias do grupo que compreende o betume, o piche, o asfalto, etc.

10. Blindagem protetora de acordo com a reivindicação 7, caracterizada pelo fato de dita pelo menos uma camada de proteção química compreender uma ou mais substâncias do grupo que compreende o breu e o colofônio.

11. Blindagem protetora de acordo com a reivindicação 8, caracterizada pelo fato de ditas substâncias inorgânicas compreenderem o enxofre.

12. Blindagem protetora de acordo com a reivindicação 8, caracterizada pelo fato de ditas substâncias inorgânicas compreenderem o cloreto de amônio.

13. Blindagem protetora de acordo com as reivindicações 5 ou 6, caracterizada pelo fato de dita camada de proteção química compreender compostos orgânicos que libertam substâncias agressivas sob a ação do calor.

14. Blindagem protetora de acordo com a reivindicação 5, caracterizada pelo fato de dita pelo menos uma camada de proteção mecânica compreender uma placa metálica (39) constituída por uma placa de base (39b) revestida por uma camada de liga de alta dureza (39a) com valor pelo menos igual a 50 HRc., aproximadamente equivalente a 500 HV1.

15. Blindagem protetora de acordo com a reivindicação 5, caracterizada pelo fato de cada dito painel compreender adicionalmente pelo menos uma placa de aço inoxidável (42).

16. Blindagem protetora de acordo com a reivindicação 15, caracterizada pelo fato de dita placa de aço inoxidável ser do tipo AISI 304.

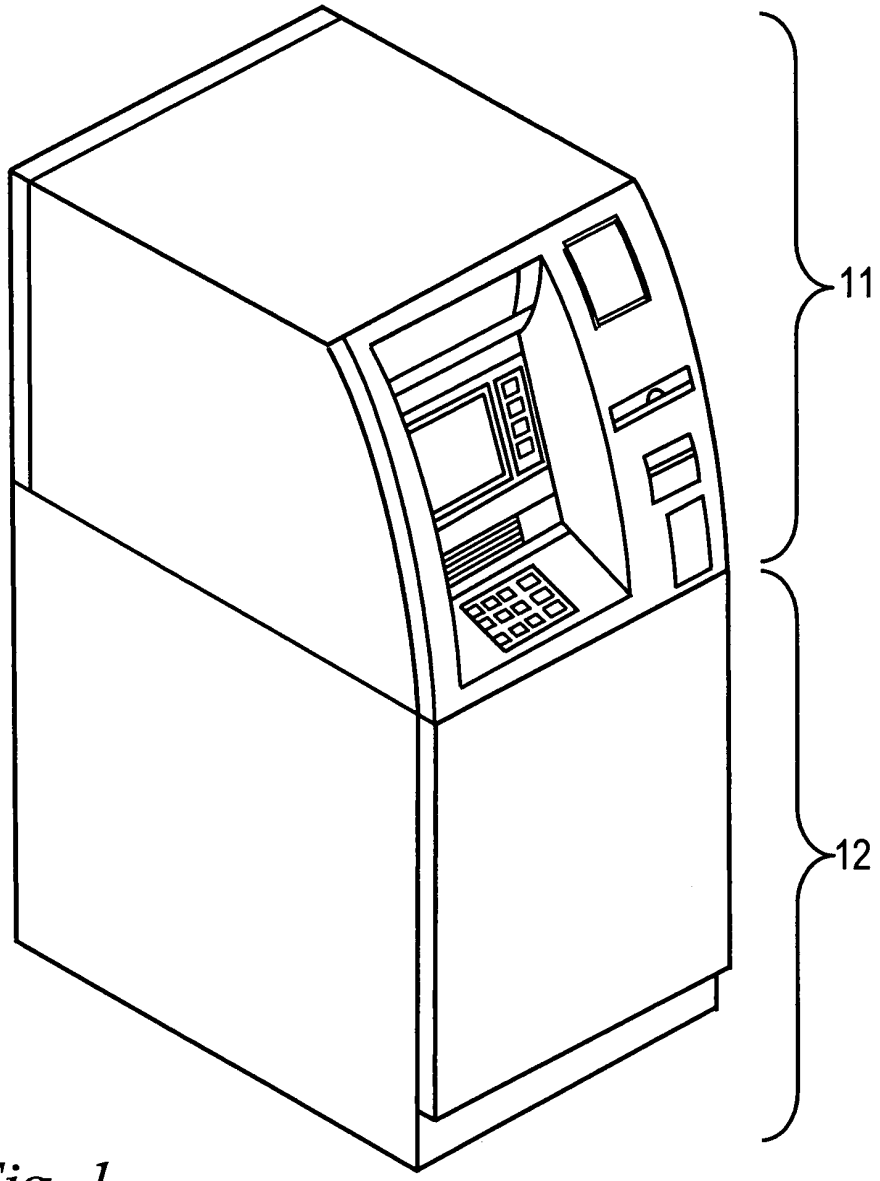


Fig. 1

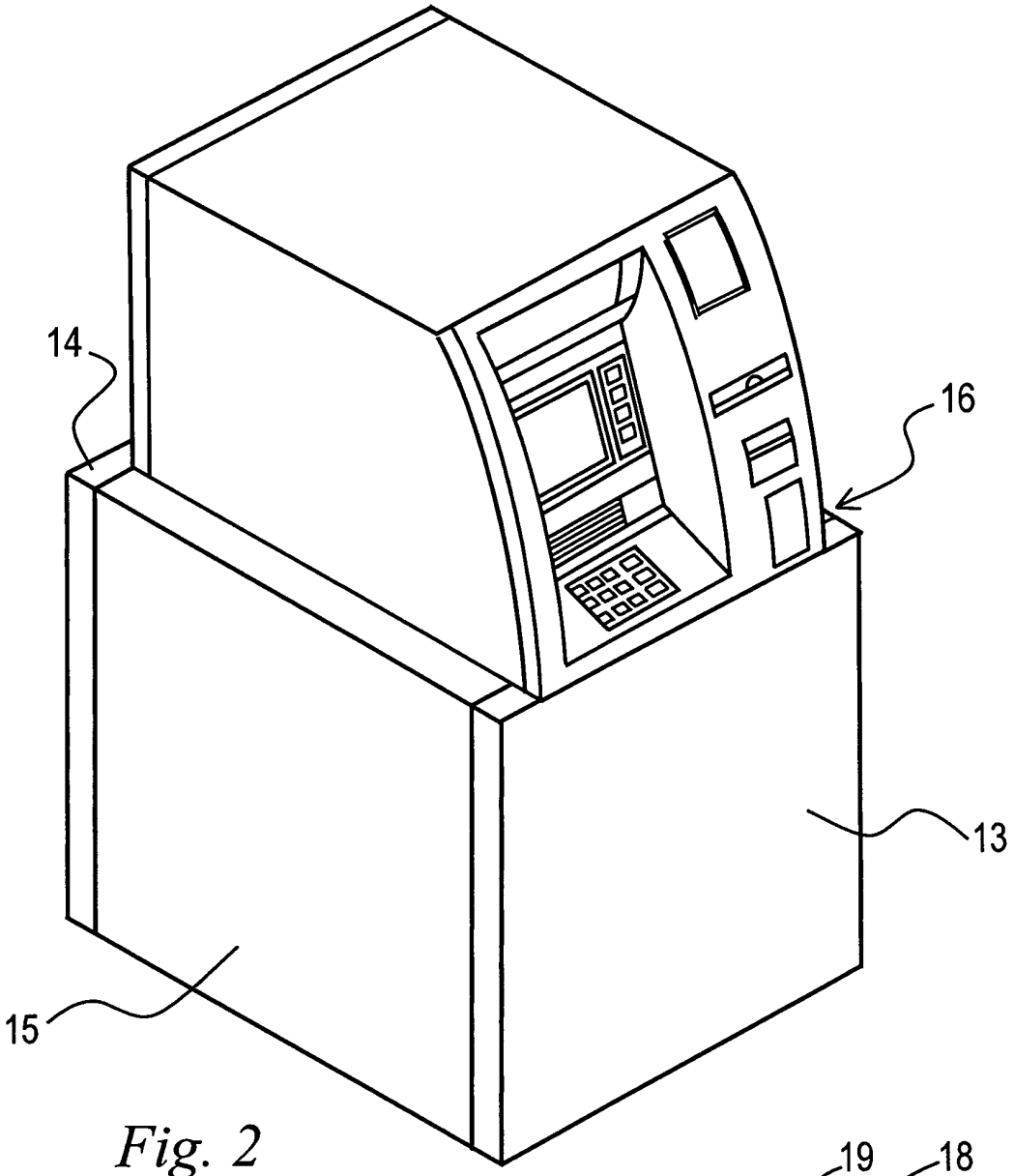


Fig. 2
(técnica anterior)

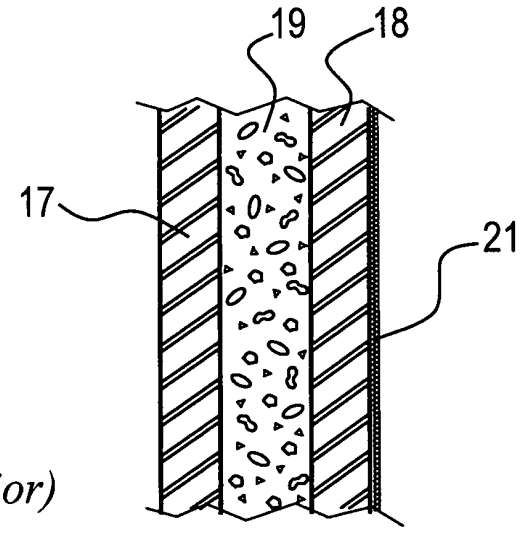


Fig. 3
(técnica anterior)

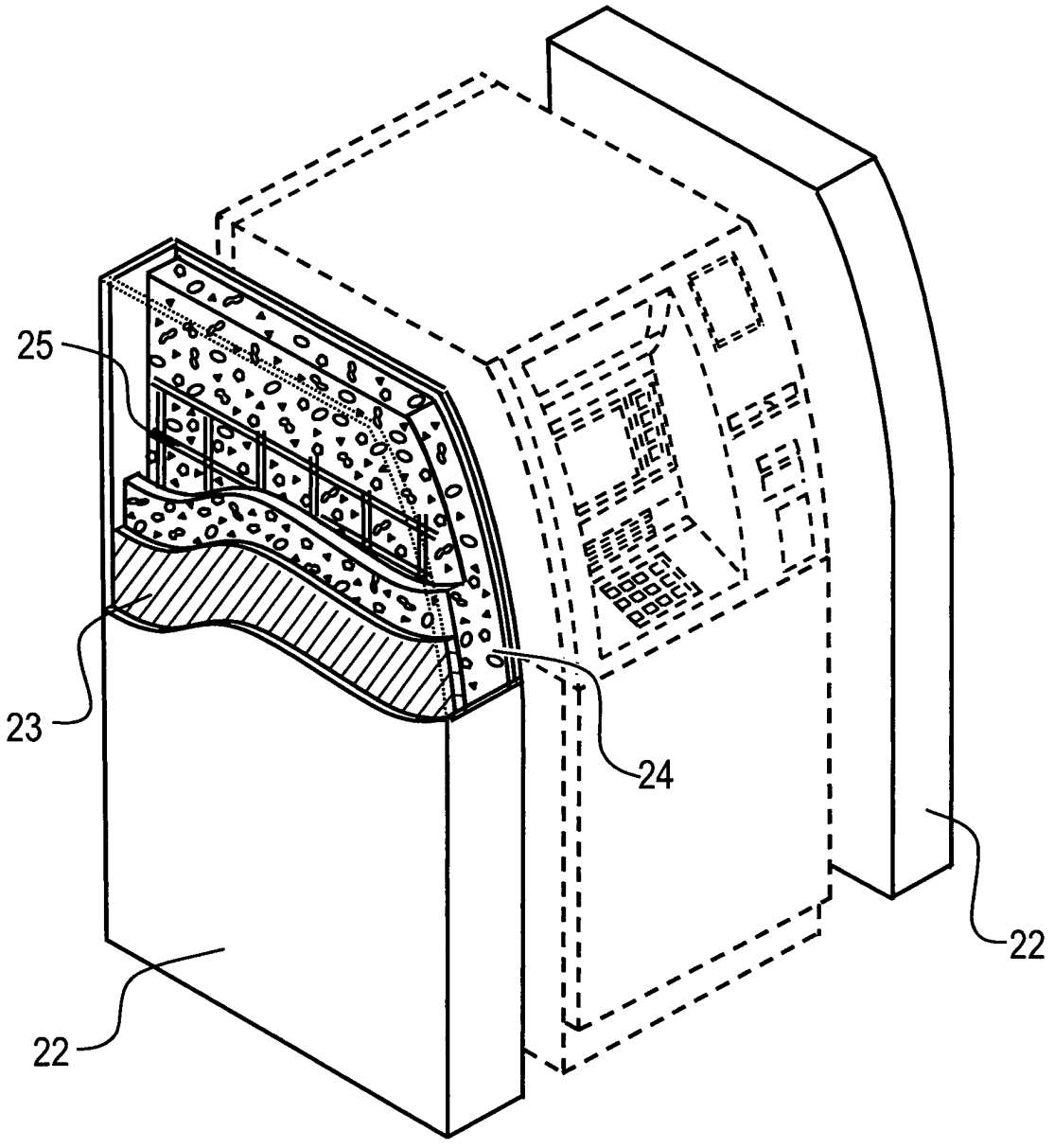


Fig. 4
(técnica anterior)

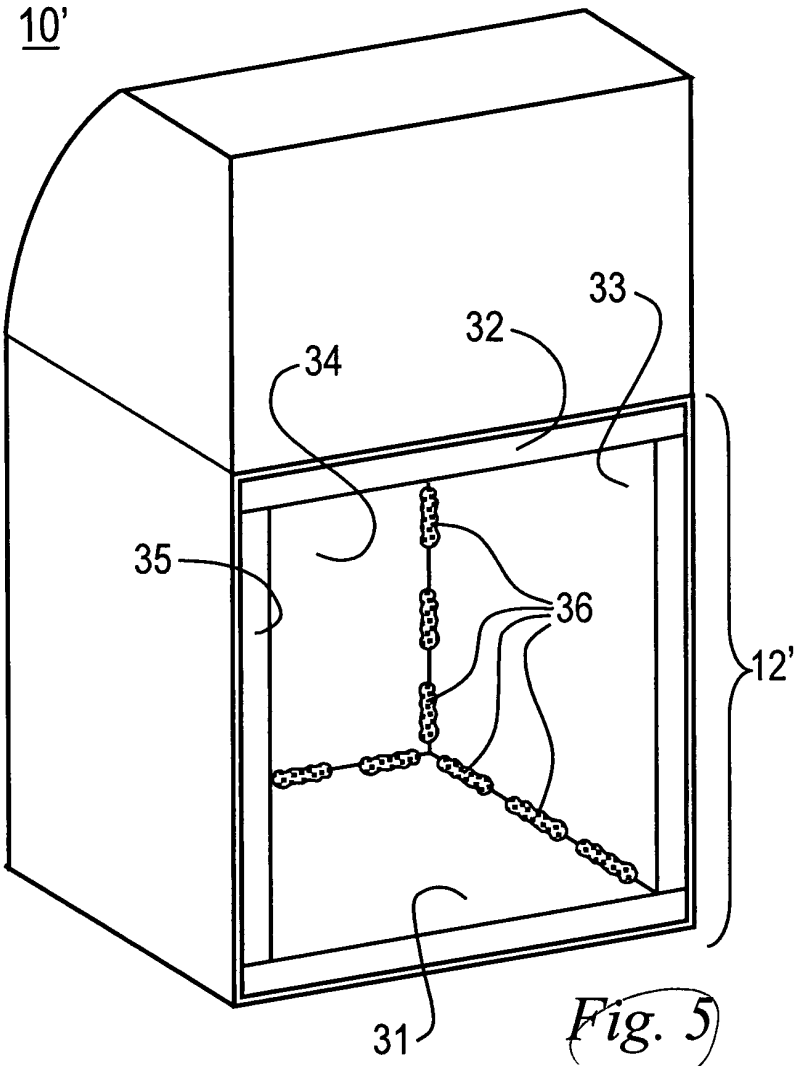


Fig. 5

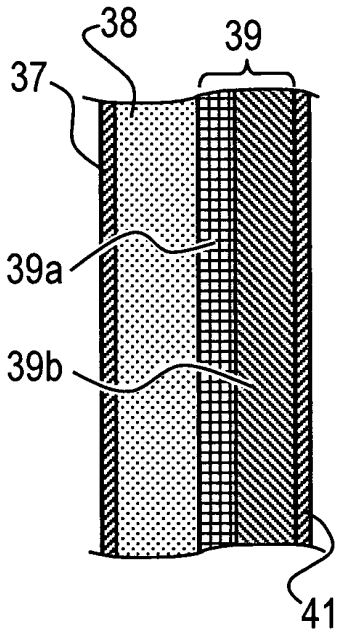


Fig. 6a

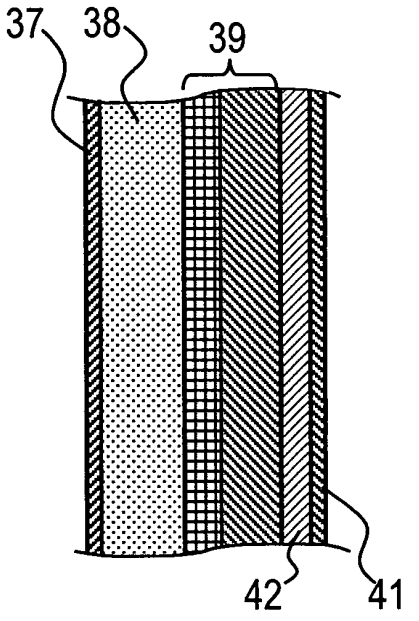


Fig. 6b

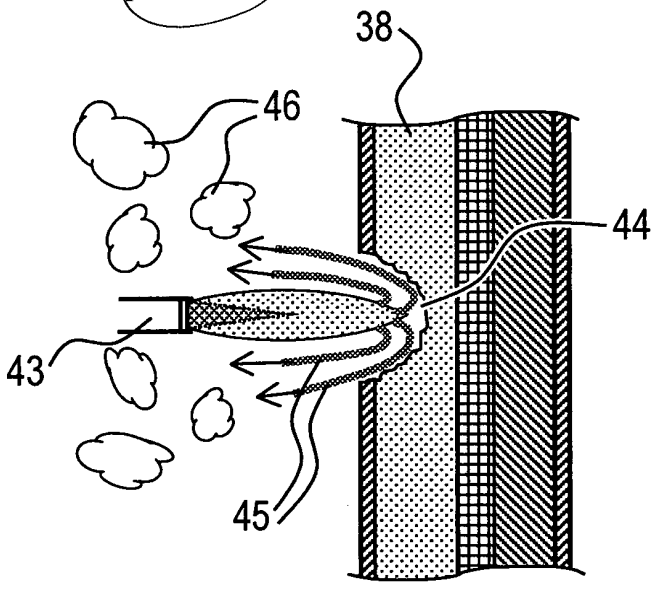


Fig. 6c

RESUMO

"BLINDAGEM PROTETORA CONTRA ARROMBAMENTO DE COFRES"

Blindagem protetora contra arrombamento de cofres compreendendo um conjunto de painéis de blindagem (31, 32, 33, 34, 35), justapostos às faces internas das paredes, piso e teto do cofre e unidos solidamente entre si ao longo de suas bordas mediante soldagem (36). Cada painel está formado por várias camadas protetoras, compreendendo proteção química (38), proteção mecânica e proteção contra perfuração (39) contra ferramentas cortantes. A proteção química compreende substâncias que, sob a ação da chama de um maçarico, desprendem compostos voláteis agressivos às mucosas do operador, enquanto a proteção contra perfuração compreende uma camada de material de elevada dureza (39a), superior a 500 HV1. Opcionalmente, cada painel pode estar provido de uma placa de aço inoxidável (42).

Apêndice D

Documentos utilizados nos experimentos e amostra dos resultados obtidos

República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) **MU 8903036-2 U2**

* B R M U 8 9 0 3 0 3 6 U 2 *

(22) Data de Depósito: 06/11/2009
(43) Data da Publicação: 18/09/2012
(RPI 2176)

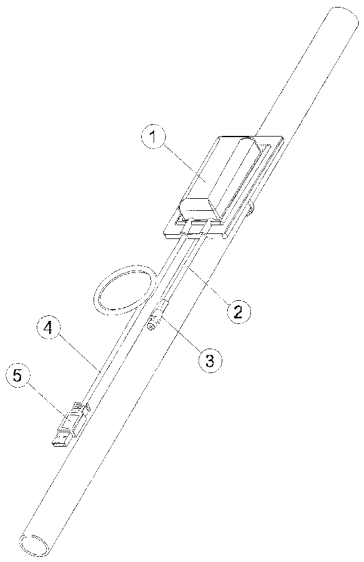
(51) Int.Cl.:
H04W 88/02
H04H 20/53

(54) Título: ESTAÇÃO EXTERNA USB

(73) Titular(es): PROQUALIT TELECOM LTDA

(72) Inventor(es): ALEXANDRE NUNES DA TRINDADE

(57) Resumo: ESTAÇÃO EXTERNA USB. (1), Patgente de Modelo de Utilidade para dispositivo independente com rádio USB, para operar em transmissão e recepção de sinal Wifi nas freqüências de 2,4 e 5,8 GHz, para uso externo com qualquer modelo de antena. Protegido contra intempéries, com saída em cabo de rede (4) e conector USB (5) na extremidade, podendo ser conectado diretamente a entrada USB de um computador e entrada em cabo coaxial (2) com conector coaxial (3), para acoplamento na antena. Possui um suporte de fixação (6) que serve para facilitar a instalação utilizando sistema de fixação (7) com a vantagem de se adaptar a qualquer tipo de antena. Ao contrátio dos convencionais a Estação Externa USB (1), possui uma placa Wireless USB (9) embutida que permite uso externo com qualquer modelo de antena, protegida contra intempéries, com saída em cabo de rede (4) e conector USB (5) na extremidade, podendo ser conectado diretamente a entrada USB de um computador e entrada em cabo coaxial (2) com conector Coaxial (3), para acoplamento na antena.



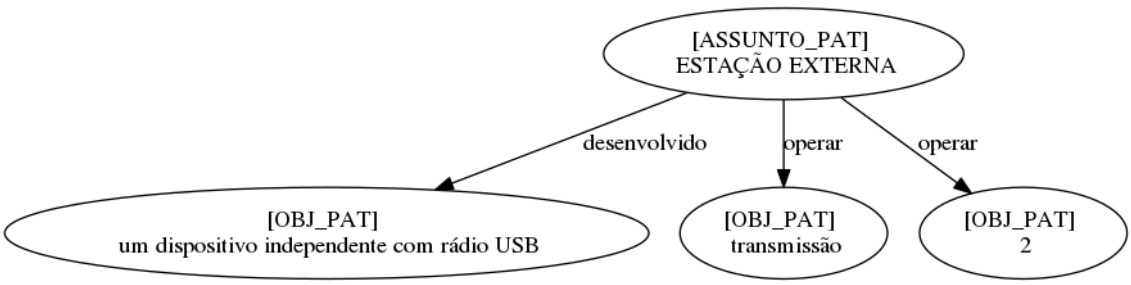
Reivindicação :

ESTAÇÃO EXTERNA USB 1, desenvolvido um dispositivo independente com rádio USB, para operar em transmissão e recepção de sinal Wifi nas frequências de 2,4 e 5,8 GHz, para uso externo com qualquer modelo de antena, protegido contra intempéries, com saída em cabo de rede 4 e conector USB 5 na extremidade, podendo ser conectado diretamente a entrada USB de um computador.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT ESTAÇÃO EXTERNA USB
    (ILUST_REF 1)
  ) ,
  (CARACT_ASSUNTO desenvolvido um
    (OBJ_PAT dispositivo independente com rádio USB)
  ) ,
  (CARACT_ASSUNTO para operar em transmissão e recepção de sinal Wifi em_ as frequências
   de 2 , 4 e 5 , 8 GHz) ,
  (CARACT_ASSUNTO para uso externo com qualquer modelo de antena) ,
  (CARACT_ASSUNTO protegido contra intempéries) ,
  (CARACT_ASSUNTO com
    (OBJ_PAT saída em cabo de rede
      (ILUST_REF 4)
    ) e
    (OBJ_PAT conector USB
      (ILUST_REF 5)
      (CARACT_OBJ em_ a extremidade)
    )
  ) ,
  (CARACT_ASSUNTO podendo ser conectado diretamente a entrada USB de um computador)
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) **PI0803602-0 A2**

* B R P I 0 8 0 3 6 0 2 A 2 *

(22) Data de Depósito: 27/06/2008
(43) Data da Publicação: 31/05/2011
(RPI 2108)

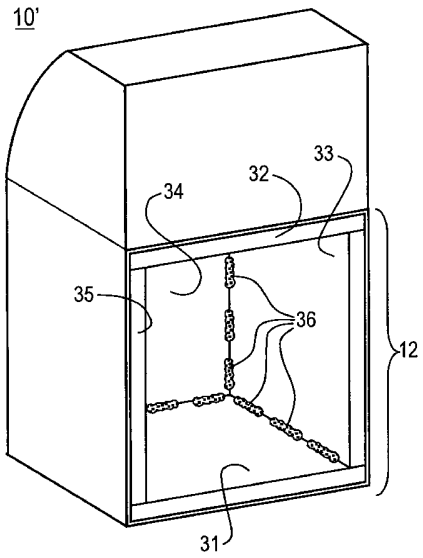
(51) Int.Cl.:
E05G 1/024 2006.01
E05G 1/10 2006.01

(54) Título: **BLINDAGEM PROTETORA CONTRA ARROMBAMENTO DE COFRES**

(73) Titular(es): ITAUTEC S.A.- GRUPO ITAUTEC

(72) Inventor(es): Ronaldo Marques, Vanderley de Assis Reis

(57) Resumo: BLINDAGEM PROTETORA CONTRA ARROMBAMENTO DE COFRES Blindagem protetora contra arrombamento de cofres compreendendo um conjunto de painéis de blindagem (31, 32, 33, 34, 35), justapostos às faces internas das paredes, piso e teto do cofre e unidos solidamente entre si ao Longo de suas bordas mediante soldagem (36). Cada painel está formado por várias camadas protetoras, compreendendo proteção química (38), proteção mecânica e proteção contra perfuração (39) contra ferramentas cortantes. A proteção química compreende substâncias que, sob a ação da chama de um maçarico, desprendem compostos voláteis agressivos às mucosas do operador, enquanto a proteção contra perfliração compreende uma camada de material de elevada dureza (39a), superior a 500 1-1V 1. Opcionalmente, cada painel pode estar provido de uma placa de aço inoxidável (42).



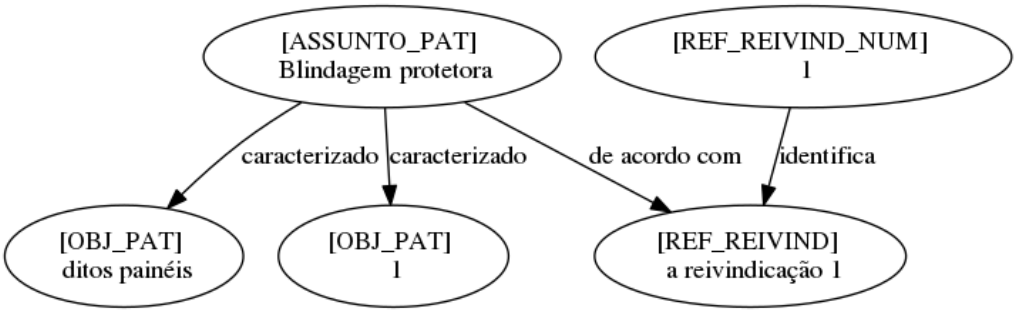
Reivindicação :

Blindagem protetora de acordo com a reivindicação 1, caracterizada pelo fato de ditos painéis estarem solidamente unidos entre si ao longo de suas bordas.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT Blindagem protetora)
  (REF de acordo com
    (REF_REIVIND a reivindicação
      (REF_REIVIND_NUM 1)
    )
  ) ,
  (CARACT_ASSUNTO caracterizada por_ o fato de
    (OBJ_PAT ditos painéis
      (CARACT_OBJ estarem solidamente unidos entre si a_ o longo de
        (OBJ_PAT suas bordas)
      )
    )
  )
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) **MU8800580-1 U2**

* B R M U 8 8 0 0 5 8 0 U 2 *

(22) Data de Depósito: 11/08/2008
(43) Data da Publicação: 11/05/2010
(RPI 2053)

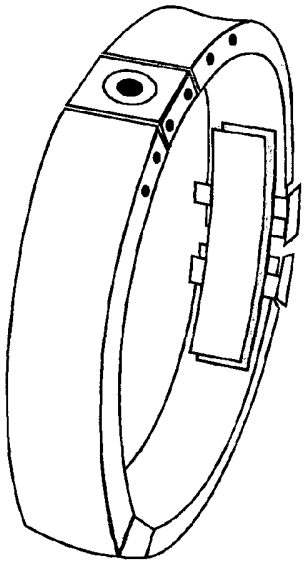
(51) Int.Cl.:
G08C 21/00 (2010.01)

(54) Título: **EQUIPAMENTO DE LOCALIZAÇÃO DE PESSOAS**

(73) Titular(es): Renato Alves de Melo

(72) Inventor(es): Renato Alves de Melo

(57) Resumo: Equipamento de Localização de Pessoas. Patente Modelo de Utilidade, elaborada para ser utilizada em lugares com grande concentração de pessoas, tais corno: shoppings, clubes, pecuárias, exposições-feiras, shows, restaurantes, etc. cuja organização se interesse na instalação deste equipamento em dar maior segurança e comodidade aos usuarios. Importante para localização de pessoas em lugares de grande aglomeração, pela facilidade de encontrar alguém no meio da multidão sem preocupação ou aborrecimento, pois permite a localização exata e instantânea da pessoa que estiver utilizando ao Equipamento de Localização de Pessoas.



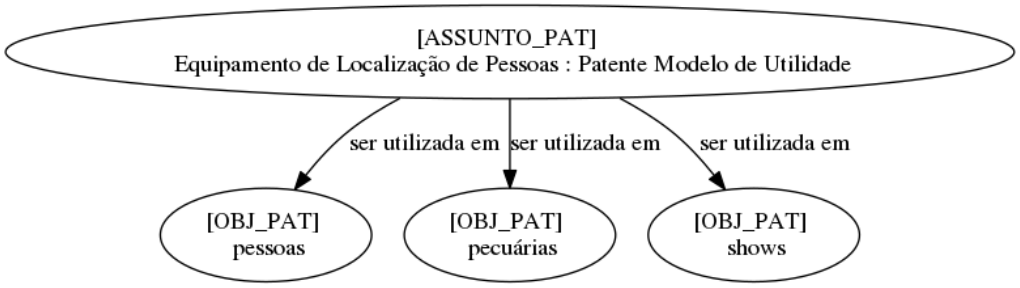
Reivindicação :

Equipamento de Localização de Pessoas: Patente Modelo de Utilidade, elaborada para ser utilizada em lugares com grande concentração de pessoas, tais como: shoppings, clubes, pecuárias, exposiçõesfeiras, shows, restaurantes, etc, cuja organização se interesse na instalação deste equipamento em dar maior segurança e comodidade aos usuários.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT Equipamento de Localização de Pessoas) :
  (CARACT_ASSUNTO Patente Modelo de Utilidade) ,
  (CARACT_ASSUNTO elaborada para ser utilizada em lugares com grande concentração
   de pessoas , tais como : shoppings , clubes , pecuárias , exposiçõesfeiras ,
   shows , restaurantes , etc , cuja organização se interesse
   em_ a instalação deste equipamento em dar maior segurança e comodidade a_ os usuários
  )
)
```

Grafo extraído :



REPÚBLICA FEDERATIVA DO BRASIL
Ministério da Indústria, do Comércio e do Turismo
Instituto Nacional da Propriedade Industrial

(11) (21) **PI 9703620-0 A**

(51) Int. Cl[5].:
B41J 27/00

(22) Data de Depósito: 18/06/1997
(43) Data de Publicação: 27/10/98 (RPI 1451)

(54) Título: **FILTRO PARA CABEÇA DE IMPRESSÃO A JATO DE TINTA**

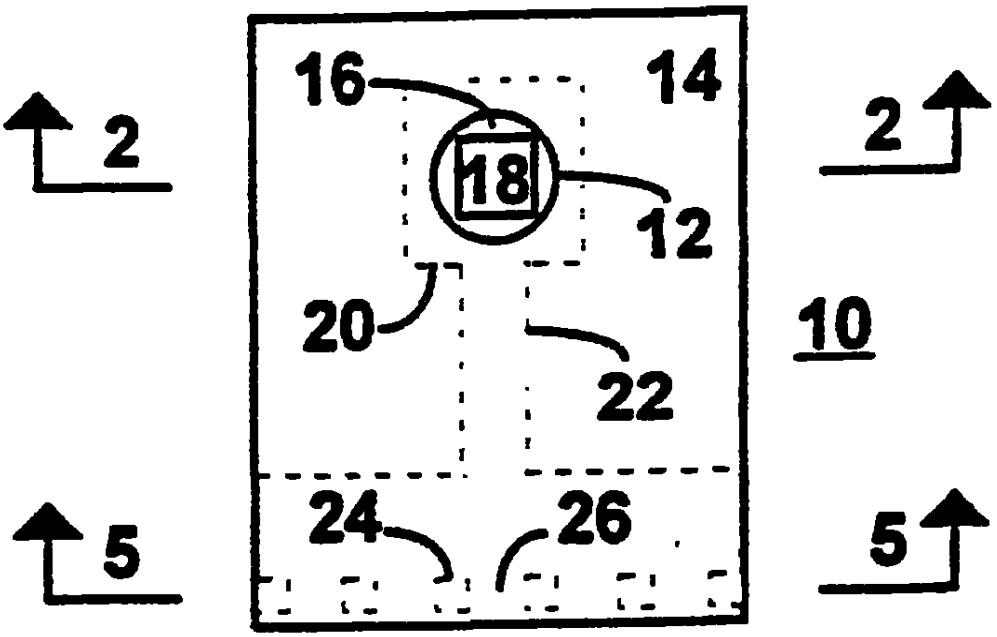
(30) Prioridade Unionista: 18/06/1996 US 665707

(71) Depositante(s): Lexmark International, Inc. (US)

(72) Inventor(es): Micah Abraham Kaufman,Janine Marie Kelty,James Harold Powers,Lawrence Russell Steward,Michael Raulinaitis

(74) Procurador: Dannemann, Siemsen, Bigler & Ipanema Moreira

(57) Resumo: Patente de Invenção: " FILTRO PARA CABEÇA DE IMPRESSÃO A JATO DE TINTA " . Uma cabeça de impressão a jato de tinta inclui uma camada de chip tendo uma via para receber tinta e tendo pelo menos um atuador. Uma camada de barreira é disposta adjacente à camada de chip, e forma pelo menos uma garganta a qual tem uma largura e uma área de seção em corte. A camada de barreira também forma pelo menos uma câmara de bolha com cada garganta adaptada para receber a tinta desde a via e provê-la para a respectiva câmara de bolha. Uma camada de injetor é disposta adjacente à camada de barreira, oposta à camada de chip, e forma pelo menos um injetor para ejetar a tinta da respectiva câmara de bolha quando a tinta é energizada pelo atuador associado. Pelo menos um coluna é disposta próximo à garganta, e se estende em parte do caminho entre a camada de chip e a camada de injetor. A coluna forma uma ou mais portas através das quais a tinta deve passar da via para a pelo menos uma garganta.



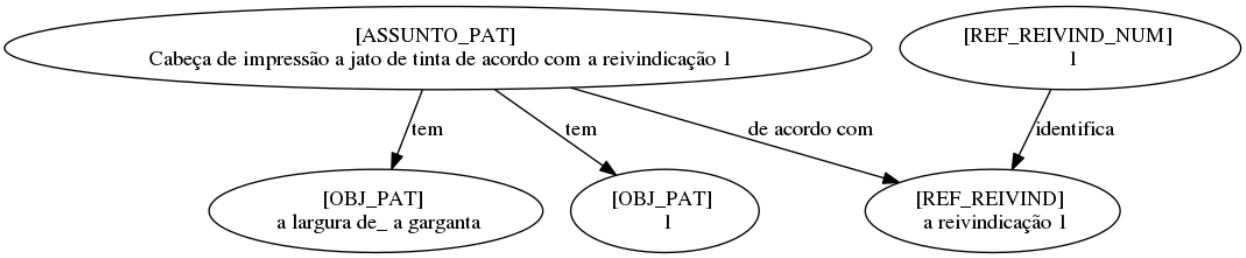
Reivindicação :

Cabeça de impressão a jato de tinta de acordo com a reivindicação 1, onde cada porta tem uma largura que é igual à largura da garganta.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT Cabeça de impressão a jato de tinta)
  (REF de acordo com
    (REF_REIVIND a reivindicação
      (REF_REIVIND_NUM 1)
    )
  ) ,
  (CARACT_ASSUNTO onde cada porta tem uma largura que é igual a_ a largura de_
    (OBJ_PAT a garganta)
  )
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) BR 10 2012 016968-1 A2

(22) Data de Depósito: 10/07/2012
(43) Data da Publicação: 06/05/2014
(RPI 2261)

(51) Int.Cl.:
B60R 25/04

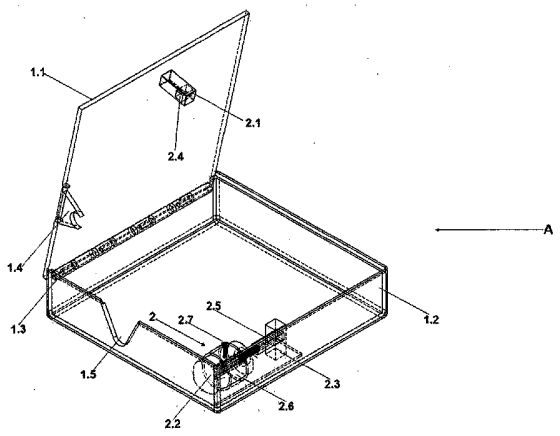
(54) Título: DISPOSITIVO ANTIFURTO VEICULAR

(66) Prioridade Interna: 860446

(73) Titular(es): André Luiz Penachio

(72) Inventor(es): André Luiz Penachio

(57) Resumo: DISPOSITIVO ANTIFURTO VEICULAR A presente invenção trata de um dispositivo antifurto veicular, compreendendo uma caixa blindada e um sistema anti-arrombamento podendo também proteger um aparelho de rastreamento veicular que possa opcionalmente ser instalado no veículo.



Reivindicação :

Dispositivo antifurto caracterizado por compreender: 1 Uma caixa blindada e, 2 Um sistema antiarrombamento.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT Dispositivo antifurto)
  (CARACT_ASSUNTO caracterizado por compreender :
    (OBJ_PAT
      (ILUST_REF 1)
      Uma caixa blindada
    ) e ,
    (OBJ_PAT
      (ILUST_REF 2)
      Um sistema antiarrombamento)
  )
)
```

Grafo extraído :

República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) MU 9001683-1 U2

BRMU9001683U2

(22) Data de Depósito: 28/09/2010
(43) Data da Publicação: 11/02/2014
(RPI 2249)

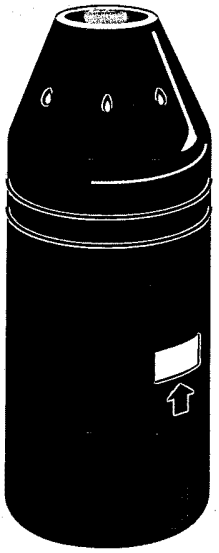
(51) Int.Cl.:
A47J 41/00

(54) Título: MEDIDOR DE CONSUMO PARA PORTA GARRAFAS

(73) Titular(es): Rosalem Souza Gois

(72) Inventor(es): ROSALEM SOUZA GOIS

(57) Resumo: 1) MEDIDOR DE CONSUMO PARA PORTA GARRAFAS. Patente de Modelo de Utilidade para recepientes térmicos para garrafas de bebidas denominado comumente de porta cerveja, acrescido do medidor de consumo. O medidor de consumo consiste de recepientes de plástico ou assemelhado, isopor ou outro material térmico (1), mola de metal ou assemelhado (2) bandeja de apoio plástico ou assemelhado (3),janelas expositoras recortadas no corpo principal da peça (4) e faixas demarcativas ou sinalizações similares (5). Cabe enfatizar que não há mudanças estrutural, nem de manuseio no processo de fechamento das tampas de fixação. Assim, como nas embalagens usuais, sua aplicação é para acondicionar e conservar a baixa temperatura garrafas de bebidas. Porém sua utilização difere do uso tradicional ao se vizualizar através da janela expositora, sem ter que abrir o recepiente, a quantidade de líquido existente nas garrafas. O medidor poderá ser confeccionado e produzido em tamanhos e quantidades diversas para atender diferentes necessidades.



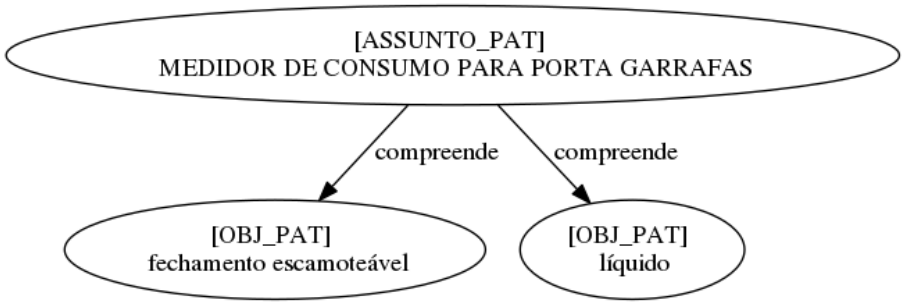
Reivindicação :

MEDIDOR DE CONSUMO PARA PORTA GARRAFAS é compreendido por embalagem com tampa de fechamento escamoteável e compartimentos formados por recipientes plásticos e de material térmico, caraterizada pelo fato da bandeja de apoio plástico ou assemelhado se mover em função da mola de metal ou assemelhado, que se expande ou retrai, demonstrando através do peso das garrafas e pelas janelas expositoras, o quanto de líquido foi consumido do interior das mesmas, sem ter que retirá las.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT MEDIDOR DE CONSUMO PARA PORTA GARRAFAS)
  (CARACT_ASSUNTO é compreendido por
    (OBJ_PAT embalagem
      (CARACT_OBJ com tampa de fechamento escamoteável)) e
    (OBJ_PAT compartimentos
      (CARACT_ASSUNTO formados por
        (OBJ_PAT recipientes plásticos e
          (CARACT_OBJ de material térmico))
      )
    )
  ) ,
  (CARACT_ASSUNTO caraterizada por_ o fato de_
    (OBJ_PAT a bandeja de apoio plástico ou assemelhado) se mover em função de_
    (OBJ_PAT a mola de metal ou assemelhado ,
      (CARACT_OBJ que se expande ou retrai) ,
      (CARACT_OBJ demonstrando através de_ o peso de_ as garrafas e por_ as
        (OBJ_PAT janelas expositoras) , o quanto de líquido foi consumido de_ o
        interior de_ as mesmas , sem ter que retirá las
      )))
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) MU 9101172-8 U2

(22) Data de Depósito: 03/06/2011
(43) Data da Publicação: 09/07/2013
(RPI 2218)

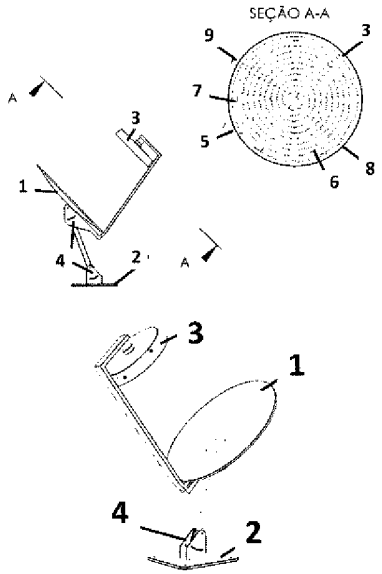
(51) Int.Cl.:
F24J 2/12

(54) Título: AQUECEDOR SOLAR DE ÁGUA QUE UTILIZA UM CONCENTRADOR SOLAR PARABÓLICO ESTACIONÁRIO

(73) Titular(es): Solar Engenharia Sustentavel LTDA

(72) Inventor(es): Alexandre Amorim Souza

(57) Resumo: AQUECEDOR SOLAR DE ÁGUA QUE UTILIZA UM CONCENTRADOR SOLAR PARABÓLICO ESTACIONÁRIO. Patente de modelo de Utilidade para um sistema de aquecimento solar de água a partir de um concentrador solar parabólico fixo 1, visando substituir os sistema atuais, e consequente reduzir os impctos ambientais e emissão dos gases - estufa. O modelo é composto por uma superfície parabólica refletiva 1 estacionária que concentra toda a radiação solar no foco da parábola onde se encontra instalado um trocador de calor 3 que é aquecido a uma temperatura muito elevada.Consequentemente a águaque circulana serpentina circular 6 é aquecida e armazenada em um reservatório térmico, que mantém a sua temperatura até o reinício do ciclo no dia seguinte, garantindo então uma temperatura agradável para banho. As características geométricas do trocador de calor 3 e da superfície parabólica refletiva 1, além da utilização da base 2 e do sistema de ajuste angular 4 possibilitam o posicionamento correto do modelo, sem que seja necessária a utilização de um sistema de rastreamento solar.



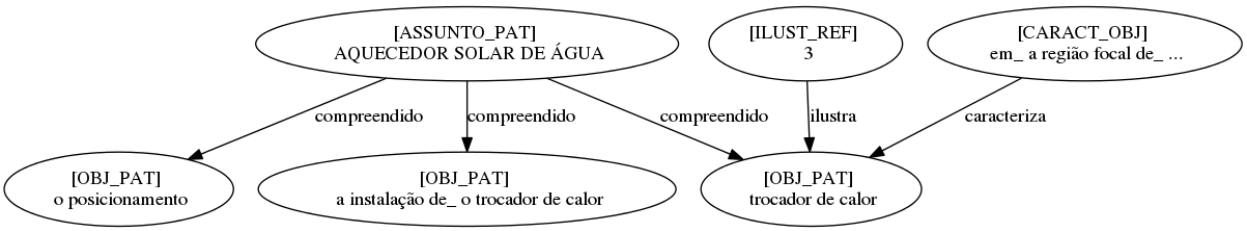
Reivindicação :

AQUECEDOR SOLAR DE ÁGUA QUE UTILIZA UM CONCENTRADOR SOLAR PARABÓLICO ESTACIONÁRIO compreendido por uma superfície parabólica refletiva 1, que utiliza a base de fixação 2 para sua instalação, o sistema de ajuste angular 4 para se obter o posicionamento adequado do modelo, que tem um braço metálico para a instalação do trocador de calor 3 na região focal da superfície parabólica refletiva 1.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT AQUECEDOR SOLAR DE ÁGUA QUE UTILIZA
   UM CONCENTRADOR SOLAR PARABÓLICO ESTACIONÁRIO)
  (CARACT_ASSUNTO compreendido por
    (OBJ_PAT uma superfície parabólica refletiva (ILUST_REF 1) ,
      (CARACT_OBJ que utiliza
        (OBJ_PAT a base de fixação (ILUST_REF 2)
          (CARACT_OBJ para sua instalação)
        )
      )
    )
  ) ,
  (CARACT_ASSUNTO
    (OBJ_PAT o sistema de ajuste angular (ILUST_REF 4)
      (CARACT_OBJ para se obter o posicionamento adequado de_
        (OBJ_PAT o modelo ,
          (CARACT_OBJ que tem
            (OBJ_PAT um braço metálico
              (CARACT_OBJ para a instalação de_
                (OBJ_PAT o trocador de calor (ILUST_REF 3)) em_ a região focal de_
                (OBJ_PAT a superfície parabólica refletiva (ILUST_REF 1)))))))))
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria
e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) **PI 1005279-8 A2**

BRPI1005279A2

(22) Data de Depósito: 29/12/2010
(43) Data da Publicação: 16/04/2013
(RPI 2206)

(51) Int.Cl.:
C12C 11/00

(54) Título: USO DE FOTO-ESTIMULAÇÃO PARA ACELERAÇÃO DA FERMENTAÇÃO NO PROCESSO DE PRODUÇÃO DE CERVEJA E CHOPE

(73) Titular(es): Vanderlei Salvador Bagnato

(72) Inventor(es): EVERTON SERGIO ESTRACANHOLLI, IGOR POLIKARPOV, VENDERLEI SALVADOR BAGNATO

(57) Resumo: USO DE FOTO-ESTIMULAÇÃO PARA ACELERAÇÃO DA FERMENTAÇÃO NO PROCESSO DE PRODUÇÃO DE CERVEJA E CHOPE. Refere-se a presente patente de invenção na utilização de fontes de luz - Laser ou LEDs ou Lâmpadas Fluorescentes, como forma de acelerar o processo de fermentação do mosto cervejeiro, que posteriormente irá se transformar em cerveja ou chope, ficando apto para o consumo, diminuindo assim o tempo de produção, diminuindo também o custo de produção e aumentando a capacidade produtiva.

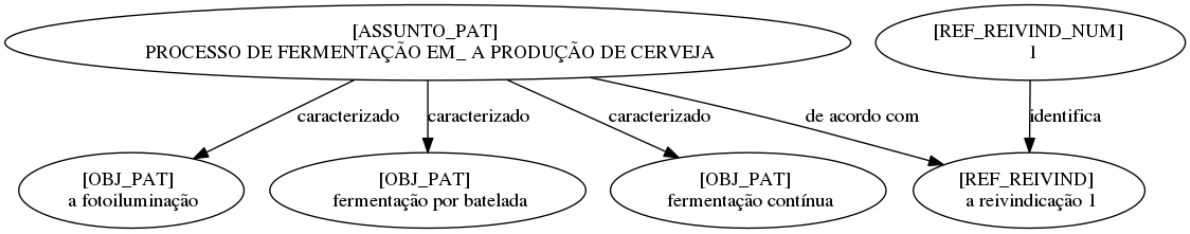
Reivindicação :

PROCESSO DE FERMENTAÇÃO NA PRODUÇÃO DE CERVEJA OU CHOPE de acordo com a reivindicação 1, caracterizado pela fotoiluminação poder ser utilizado tanto no processo de fermentação por batelada como no de fermentação contínua.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT PROCESSO DE FERMENTAÇÃO EM_ A PRODUÇÃO DE CERVEJA OU CHOPE)
  (REF de acordo com a
    (REF_REIVIND reivindicação
      (REF_REIVIND_NUM 1)
    )
  ) ,
  (CARACT_ASSUNTO caracterizado por_
    (OBJ_PAT a fotoiluminação)
    poder ser utilizado tanto em_ o processo de fermentação por batelada
    como em_ o de fermentação contínua
  )
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) MU 9101620-7 U2

* B R M U 9 1 0 1 6 2 0 U 2 *

(22) Data de Depósito: 13/07/2011
(43) Data da Publicação: 16/07/2013
(RPI 2219)

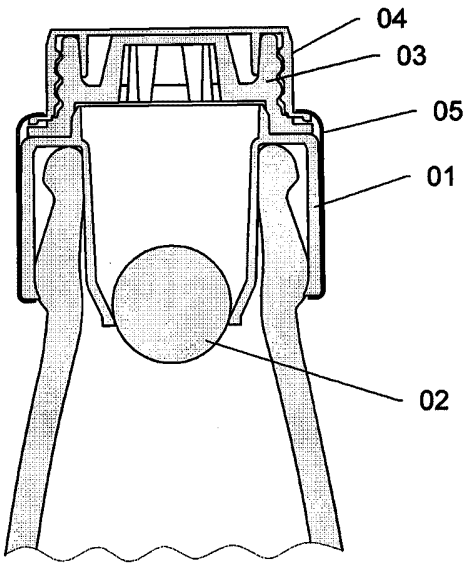
(51) Int.Cl.:
B65D 41/48
B65D 41/40
B65D 41/58
B65D 45/02

(54) Título: TAMPA INVIOLÁVEL PARA GARRAFA DE CERVEJA

(73) Titular(es): ISA Industria de Embalagens Ltda

(72) Inventor(es): Esio Missiato Junior

(57) Resumo: TAMPA INVIOLÁVEL PARA GARRAFA DE CERVEJA, composta por uma base (01) que é encaixada sobre pressão na boca da garrafa, uma esfera (02) que impede a adulteração do produto engarrafado, um vertedor (03) para facilitar o escoamento do produto, uma tampa do vertedor (04) que é rosqueada no vertedor e uma capa de alumínio (05) que envolve todo o conjunto.



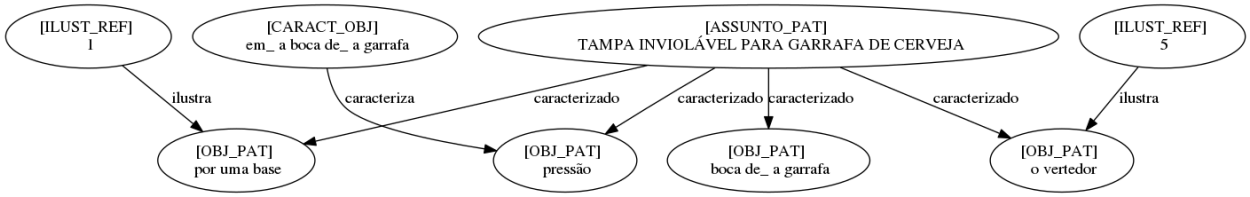
Reivindicação :

TAMPA INVIOLÁVEL PARA GARRAFA DE CERVEJA caracterizada por uma base 01 que é encaixada sobre pressão na boca da garrafa, uma esfera 02 que impede a adulteração do produto engarrafado, um vertedor 03 para facilitar o escoamento do produto, uma tampa do vertedor 04 que é rosqueada no vertedor e uma capa de alumínio 05 que envolve todo o conjunto.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT TAMPA INVIOLÁVEL PARA GARRAFA DE CERVEJA)
  (CARACT_ASSUNTO caracterizada por
    (OBJ_PAT uma base (ILUST_REF 01)
      (CARACT_OBJ que é encaixada sobre pressão em_
        (OBJ_PAT a boca de_ a garrafa)
      )
    ) ,
    (OBJ_PAT uma esfera (ILUST_REF 02)
      (CARACT_OBJ que impede a adulteração de_ o produto engarrafado)
    ) ,
    (OBJ_PAT um vertedor (ILUST_REF 03)
      (CARACT_OBJ para facilitar o escoamento de_ o produto)
    ) ,
    (OBJ_PAT uma tampa de_ o vertedor (ILUST_REF 04)
      (CARACT_OBJ que é rosqueada em_
        (OBJ_PAT o vertedor)
      )
    ) e
    (OBJ_PAT uma capa de alumínio (ILUST_REF 05)
      (CARACT_OBJ que envolve todo o conjunto)
    )
  )
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) **PI 1004834-0 A2**

BRPI1004834A2

(22) Data de Depósito: 05/11/2010
(43) Data da Publicação: 26/02/2013
(RPI 2199)

(51) Int.Cl.:
C12C 5/02
C12C 11/00

(54) Título: CERVEJA A BASE DE TRIGO COM GUARANÁ

(73) Titular(es): ON TRADE DISTRIBUIDORA DE BEBIDAS LTDA.

(72) Inventor(es): GUSTAVO NOGUEIRA SANCHES

(57) Resumo: CERVEJA A BASE DE TRIGO COM GUARANÁ. Que se fundamenta em uma formulação para 10001 a partir de: - Malte Pilsen - 115Kg - Malte de trigo Claro - 115Kg - Lúpulo hallertau tradition - 51 Og - Guaraná em pó - 750g - Fermento Welhenstephan 3068, wyeast Na brassagem deve-se arriar a 55°, manter em 53° por 30 minutos, elevar para 66° em 13 minutos e manter em 66° por 80 minutos,elevar a 78°, manter por 10 minutos e proceder à lavagem. A fermentação deve ocorrer a 19° pelos cinco primeiros dias, com pressão liberada até 5,5 platos, após elevar para 20° até 3,2 platos, com pressão fechada.

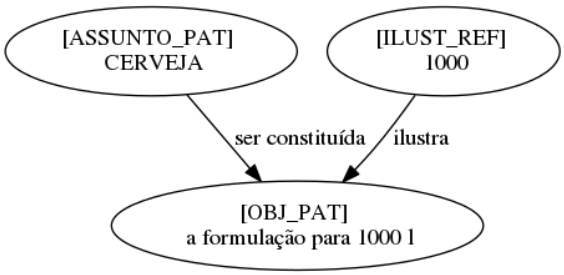
Reivindicação :

CERVEJA A BASE DE TRIGO COM GUARANÁ, caracterizado por ser constituída da formulação para 1000 l a partir de: Malte Pilsen 115Kg, Malte de trigo Claro 115Kg, Lúpulo hallertau tradition 510g, Guaraná em pó 750g, Fermento Welhenstephan 3068, wyeast.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT CERVEJA A BASE DE TRIGO COM GUARANÁ) ,
  (CARACT_ASSUNTO caracterizado por ser constituída de_
    (OBJ_PAT a formulação para 1000 l
      (CARACT_OBJ a partir de :
        (OBJ_PAT Malte Pilsen  115Kg) ,
        (OBJ_PAT Malte de trigo Claro  115Kg) ,
        (OBJ_PAT Lúpulo hallertau tradition  510g) ,
        (OBJ_PAT Guaraná em pó  750g) ,
        (OBJ_PAT Fermento Welhenstephan 3068 , wyeast)
      )
    )
  )
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial.

(21) **PI 1101244-7 A2**

* B R P I 1 1 0 1 2 4 4 A 2 *

(22) Data de Depósito: 22/03/2011
(43) Data da Publicação: 15/01/2013
(RPI 2193)

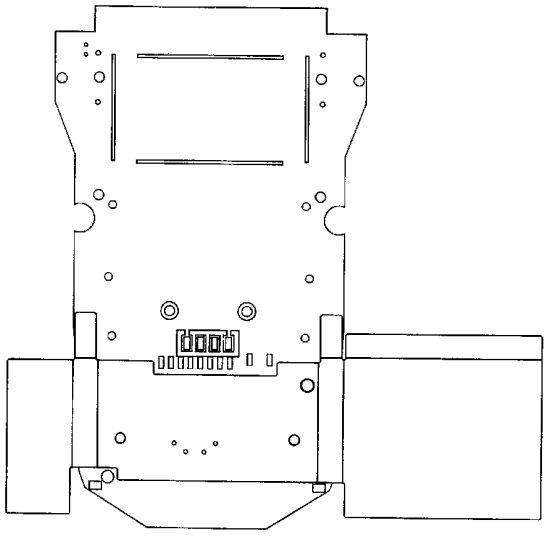
(51) Int.Cl.:
G06K 7/00
G11B 5/10
G11B 5/40
H05K 5/00

(54) Título: SISTEMA PARA PROTEÇÃO DE CONECTOR PARA CARTÕES INTELIGENTES EM EQUIPAMENTOS QUE EXIGEM SEGURANÇA DE DADOS

(73) Titular(es): TECVAN INFORMÁTICA LTDA.

(72) Inventor(es): JORGE RIBEIRO PEREIRA

(57) Resumo: SISTEMA PARA PROTEÇÃO DE CONECTOR PARA CARTÕES INTELIGENTES EM EQUIPAMENTOS QUE EXIGEM SEGURANÇA DE DADOS. Apresenta um sistema que protege as partes internas e terminais externos de Conector para Cartão Inteligente (SmartCard) com processo de envolvimento físico com cicuito impresso maleável (FPC) conectado à Placa de Circuito Impresso (PCI) com sensoriamnto. A proteção é em todas as direções, exceto na face frontal do Conector para Cartão Inteligente (SmartCard) para possibilitar a inserção do Cartão Inteligente (SmartCard) ao equipamento. O Circuito Impresso Maleável (FPC) envolve o Conector para Cartão Inteligente (SmartCard), sendo que o suporte de proteção é sobreposto, total ou parcialmente, sobre o Conector para Cartão Inteligente (SmartCard) impedindo que o FPC seja retirado. No suporte de proteção há um alojamento conector tipo Zebra que é reposável pelo sensoriamento entre as PCIs presentes nas faces superior e inferior do Conector para Cartão Inteligente (SmartCard) e também impede acesso aos terminais traseiros do conector que está sendo protegido. Na ocorrência de uma invasão ao Conector para Cartão Inteligente (SmartCard), o acionamento de qualquer um dos sensores comunica ao processador que destrói todas as chaves de segurança que criptografam as informações sigilosas armazenadas em uma memória eletrônica.



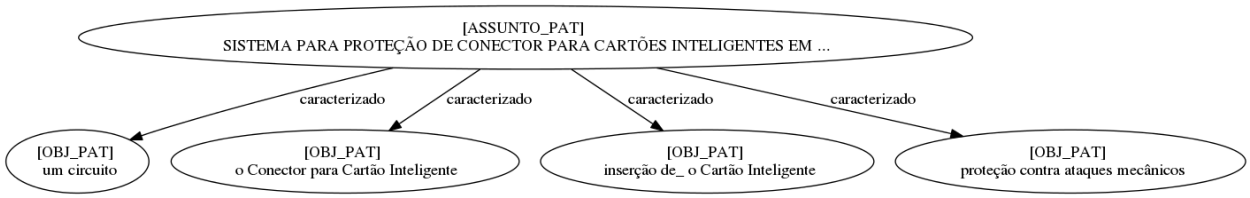
Reivindicação :

SISTEMA PARA PROTEÇÃO DE CONECTOR PARA CARTÕES INTELIGENTES EM EQUIPAMENTOS QUE EXIGEM SEGURANÇA DE DADOS caracterizado por um ou mais circuitos impressos maleáveis FPCs montados de forma a envolver todas as faces do Conector para Cartão Inteligente SmartCard, exceto sua face frontal que permanece acessível para inserção do Cartão Inteligente SmartCard no conector, que abriga um circuito interno de proteção contra ataques mecânicos, elétricos ou eletrônicos, podendo ser protegidos por adesivo ou resina, que deve ser conectado a um circuito de monitoramento de segurança.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT SISTEMA PARA PROTEÇÃO DE CONECTOR PARA CARTÕES INTELIGENTES
   EM EQUIPAMENTOS QUE EXIGEM SEGURANÇA DE DADOS)
  (CARACT_ASSUNTO caracterizado por um ou mais
    (OBJ_PAT circuitos impressos maleáveis FPCs
      (CARACT_OBJ montados de forma a envolver todas as faces de_
        (OBJ_PAT o Conector para Cartão Inteligente SmartCard)
      )  ,
      (CARACT_OBJ exceto sua
        (OBJ_PAT face frontal
          (CARACT_OBJ que permanece acessível para inserção de_ o
           Cartão Inteligente SmartCard em_
            (OBJ_PAT o conector ,
              (CARACT_OBJ que abriga
                (OBJ_PAT um circuito interno de proteção contra ataques
                 mecânicos , elétricos ou eletrônicos)))))) ,
      (CARACT_OBJ podendo ser protegidos por
        (OBJ_PAT adesivo) ou (OBJ_PAT resina)
      ) ,
      (CARACT_OBJ que deve ser conectado a um circuito de monitoramento de segurança)
    )
  )
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(11) (21) **MU 8700449-6 U**

(22) Data de Depósito: 02/01/2007
(43) Data de Publicação: **19/08/2008**
(RPI 1963)

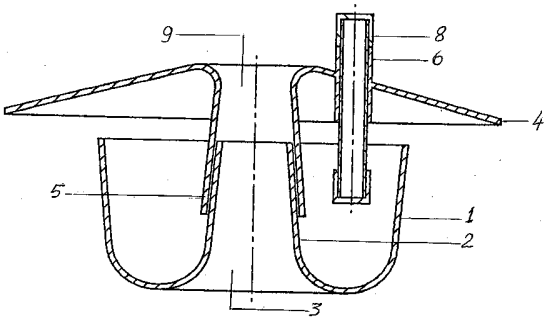
(51) Int. Cl.:
A01M 1/10 (2008.04)
A01G 9/02 (2008.04)

(54) Título: **ARANDELA PROTETORA**

(71) Depositante(s): Worny Conceição Beal (BR/RS)

(72) Inventor(es): Worny Conceição Beal

(57) Resumo: "ARANDELA PROTETORA". Patente de Modelo de Utilidade para uma arandela de proteção que é compreendida por um reservatório de forma redonda 1 com uma aba externa vertical levemente inclinada, uma projeção cônica para cima com uma abertura 2. A tampa obedece o formato redondo conforme o reservatório com uma inclinação para baixo, tendo no centro uma projeção cônica, como um tronco de cone 5 que interliga tampa ao reservatório para afastar ou fixar os mesmos, atravessando a tampa tem um apêndice 6 que serve de guia para a bóia 7 indicadora de nível de líquido.



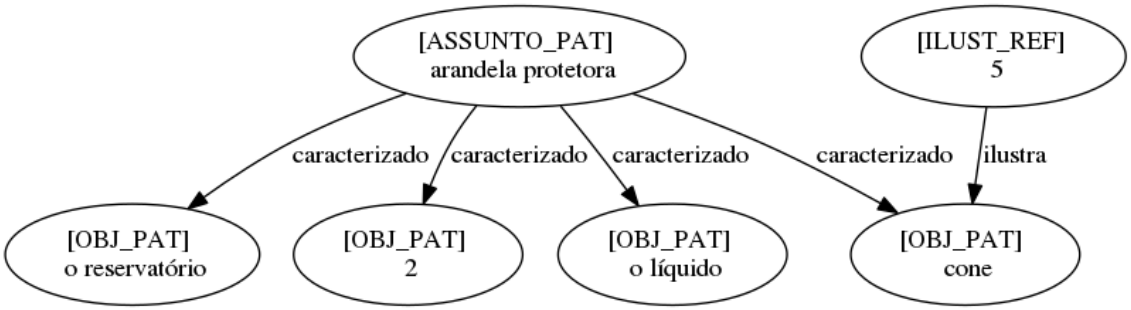
Reivindicação :

A arandela protetora, caracterizada por ter uma tampa de formato redonda 4 inclinada para baixo, tendo no centro uma projeção cônica como um tronco de cone 5 que se encaixa. no cone do reservatório 2 atravessando a mesma tem um apêndice 6 que serve de guia para a bóia indicadora de nível do liquido 7.

Anotação manual :

```
(ROOT A
  (ASSUNTO_PAT arandela protetora),
    (CARACT_ASSUNTO caracterizada por ter
      (OBJ_PAT uma tampa
        (CARACT_OBJ de formato redonda
          (ILUST_REF 4) inclinada para baixo, tendo em_ o centro
          (OBJ_PAT uma projeção cônica
            (CARACT_OBJ como um tronco de cone
              (ILUST_REF 5) que se encaixa em_ o cone de_ o reservatório (ILUST_REF 2)
            )
            (CARACT_OBJ atravessando a mesma tem
              (OBJ_PAT um apêndice
                (ILUST_REF 6)
                (CARACT_OBJ que serve de guia para
                  (OBJ_PAT a bóia indicadora de nível de_ o liquido (ILUST_REF 7) ))))))))
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) BR 10 2012 001397-5 A2

(22) Data de Depósito: 20/01/2012
(43) Data da Publicação: 21/01/2014
(RPI 2246)

(51) Int.Cl.:
A01N 63/02

(54) Título: MÉTODO PARA PRODUÇÃO DE INSETICIDAS A BASE DE B. THURINGIENSIS (BT)

(30) Prioridade Unionista: 21/01/2011 MY PI 2011000307

(73) Titular(es): Malaysian Palm Oil Board

(72) Inventor(es): Mohamed Mazmira Mohd Masri, Mohd Basri Wahid, Mohd Najib Ahmad, Siti Ramlah Ahmad Ali

(57) Resumo: MÉTODO PARA PRODUÇÃO DE INSETICIDAS À BASE DE B.thuringiensis (Bt), A presente invenção diz respeito a um método para a produção de inseticidas à base de Bacilius thuringiensis (Bt), usando o processo de fermentação. O citado método envolve o suprimento de um meio de cultura otimizado, contendo a relação carbono-nitrogênio de cerca de 2:1 (por peso) para o desenvolvimento do microorganismo que produz inseticidas à base de Bt em condições aeróbicas, sendo que o nível de oxigênio é mantido constante a aproximadamente 80%. Subsequentemente, a biomassa resultante é recuperada através do evaporador a vácuo onde a sua concentração pode ser aumentada por cerca de 20% antes de ser submetida ao processo de formulação. No processo de formulação, tais aditivos como um dispersante, diluentes e um agente de suspensão, são moídos junto com a biomassa concentrada a fim de aumentar a estabilidade dos produtos de fermentação. Figura 1 é a mais ilustrativa.

Reivindicação :

MÉTODO PARA PRODUÇÃO DE INSETICIDAS À BASE DE B. Thuringiensis Bt, caracterizado por compreender as seguintes etapas:

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT MÉTODO PARA PRODUÇÃO DE INSETICIDAS À BASE DE B. Thuringiensis Bt),
  (CARACT_ASSUNTO caracterizado por compreender as seguintes etapas:)
)
```

Grafo extraído :

República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) BR 10 2012 006350-6 A2

(22) Data de Depósito: 21/03/2012
(43) Data da Publicação: 01/10/2013
(RPI 2230)

(51) Int.Cl.:
A01M 7/00

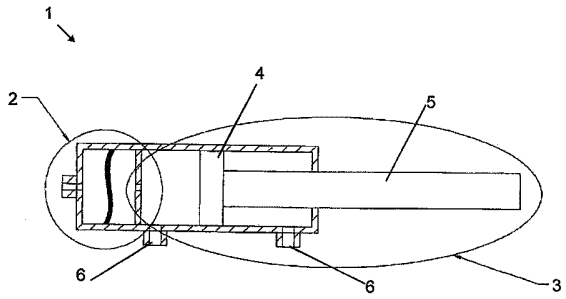
(54) Título: PISTÃO HIDRÁULICO COM ACUMULADOR INTEGRADO APLICÁVEL A BARRAS DE PULVERIZADORES AGRÍCOLAS

(30) Prioridade Unionista: 15/12/2011 BR BR102012000549-2

(73) Titular(es): Agco do Brasil Comercio e Industria LTDA.

(72) Inventor(es): Giuliano Ransolin, Luiz Gustavo Garcia

(57) Resumo: PISTÃO HIDRÁULICO COM ACUMULADOR INTEGRADO APLICÁVEL A BARRAS DE PULVERIZADORES AGRÍCOLAS. A PRESENTE INVENÇÃO REFERE-SE A UM PISTÃO HIDRÁULICO QUE COMPREENDE UM ACUMULADOR HIDRÁULICO (2) E UM ATUADOR (3) INTEGRADOS EM UMA MESMA PEÇA.
O PISTÃO HIDRÁULICO COM ACUMULADOR INTEGRADO (1) É APLICÁVEL A ASAS (13) DE PULVERIZADORES AGRÍCOLAS E É CAPAZ DE TRAZER DIVERSAS VANTAGENS AO FUNCIONAMENTO DOS PULVERIZADORES, TAIS COMO: REDUZIR A POSSIBILIDADE DE VAZAMENTOS E FALHAS DURANTE A OPERAÇÃO DO PULVERIZADOR E REDUZIR OS CUSTOS DE FABRICAÇÃO RELACIONADOS AO NÚMERO DE PEÇAS E MONTAGEM DO PULVERIZADOR.



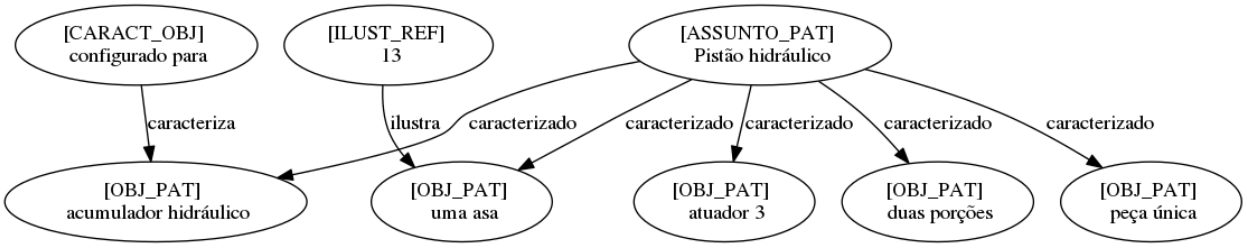
Reivindicação :

Pistão hidráulico 1 acionado através do fluido hidráulico e aplicável a uma asa 13 de um pulverizador agrícola, caracterizado pelo fato de que o pistão hidráulico 1 compreende duas porções distintas que são integradas em peça única, sendo elas: um acumulador hidráulico 2, e um atuador 3, sendo o acumulador hidráulico 2 configurado para absorver picos de pressão existentes no fluido hidráulico do interior do pistão hidráulico e o atuador 3 configurado para viabilizar a movimentação de uma haste 5 em seu interior, mediante o deslocamento do fluido hidráulico internamente ao pistão hidráulico 1.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT Pistão hidráulico (ILUST_REF 1))
  (CARACT_ASSUNTO acionado através de_ o fluido hidráulico e aplicável a
    (OBJ_PAT uma asa (ILUST_REF 13) de um pulverizador agrícola)
  ),
  (CARACT_ASSUNTO caracterizado por_ o fato de que o
    (OBJ_PAT pistão hidráulico (ILUST_REF 1)
      (CARACT_OBJ compreende duas porções distintas que são integradas em peça única,
       sendo elas:
        (OBJ_PAT um acumulador hidráulico (ILUST_REF 2)),  e
        (OBJ_PAT um atuador (ILUST_REF 3)), sendo o
        (OBJ_PAT acumulador hidráulico (ILUST_REF 2)
          (CARACT_OBJ configurado para absorver picos de pressão existentes em_
           o fluido hidráulico de_ o interior de_ o pistão hidráulico e
            (OBJ_PAT o atuador (ILUST_REF 3)
              (CARACT_OBJ configurado para viabilizar a movimentação de
                (OBJ_PAT uma haste (ILUST_REF 5)) em seu interior, mediante o
                 deslocamento de_ o fluido hidráulico internamente a_ o
                (OBJ_PAT pistão hidráulico (ILUST_REF 1)))))))))
)
```

Grafo extraído :



República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(21) BR 10 2012 032807-0 A2

(22) Data de Depósito: 20/12/2012
(43) Data da Publicação: 19/11/2013
(RPI 2237)

(51) Int.Cl.:
B01D 3/14
B01D 3/40
C12F 3/06
C12P 7/10

(54) Título: SISTEMA E PROCESSO PARA A PRODUÇÃO INTEGRADA DE ETANOL DE PRIMEIRA E SEGUNDA GERAÇÕES, E , USO DE PONTOS DE INTEGRAÇÃO PARA DITA PRODUÇÃO

(30) Prioridade Unionista: 30/03/2012 BR BR 10 2012 0072990 8

(73) Titular(es): CTC - Centro de Tecnologia Canavieira S.A.

(72) Inventor(es): Célia Maria Araújo Galvão, Danilo Ribeiro de Lima, José Augusto Travassos Rios Tomé, José Ricardo Medeiros Pinto, Juliana Conceição Teodoro, Liliane Pires Andrade, Oswald Gogoy Neto

(57) Resumo: SISTEMA E PROCESSO PARA A PRODUÇÃO INTEGRADA DE ETANOL DE PRIMEIRA E SEGUNDA GERAÇÕES, E , USO DE PONTOS DE INTEGRAÇÃO PARA DITA PRODUÇÃO. A presente invenção se refere a um sistema e um processo para a produção de etanol e produtos afins a partir de biomassas lignocelulósicas (etanol de segunda geração - 2G), em especial bagaço e palha de cana-de-açucar, mas não limitado a elas, integrado a processos convencionais de produção de etanol (etanol de primeira geração - 1G) como, por exemplo, a partir de caldo e/ou melaço de cana (processo tipicamente brasileiro, seja em usinas de açucar e etanol ou destilarias autônomas), milho,cereal,trigo,sorgo sacarino,beterraba branca, dentre outros, compreendendo reaproveitamento de correntes e efluentes. Mais especifamente, esta invenção se refere a um processo integrado para produção de etanol e produtos afins com aumento de eficiência no uso da matéria-prima, vapor, energia elétrica e água tratada, principalmente.

Reivindicação :

Sistema para produção integrada de etanol de primeira e segunda gerações, caracterizado pelo fato de compreender os seguintes pontos de integração:

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT Sistema para produção integrada de etanol de primeira e segunda gerações),
  (CARACT_ASSUNTO caracterizado por_ o fato de compreender os seguintes pontos de integração:)
)
```

Grafo extraído :

República Federativa do Brasil
Ministério do Desenvolvimento, Indústria e do Comércio Exterior
Instituto Nacional da Propriedade Industrial

(11) (21) **PI 0701138-5 A**

(22) Data de Depósito: 02/01/2007
(43) Data de Publicação: **19/08/2008**
(RPI 1963)

(51) Int. Cl.:
G07D 11/00 (2008.04)

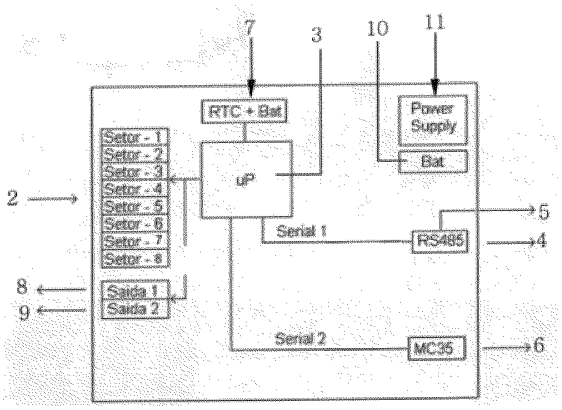
(54) Título: **SISTEMA DE ALARME PARA TERMINAIS DE AUTO-ATENDIMENTO BANCÁRIO**

(71) Depositante(s): Instalarme Industria e Comercio Ltda (BR/SP)

(72) Inventor(es): Jaime Jose Hartmann

(74) Procurador: Marpa Cons. e Asses. Empres. LTDA

(57) Resumo: "SISTEMA DE ALARME PARA TERMINAIS DE AUTO-ATENDIMENTO BANCÁRIO". Consistindo-se em um sistema de alarme especialmente desenvolvido para ser instalado dentro de terminais de auto-atendimento bancário (ATMs). Devido ao seu baixo perfil e ao seu tamanho reduzido podem ser facilmente fixados no interior dos terminais, sem prejudicar a sua operação normal de abastecimento. O sistema conta com setores disponíveis para a instalação de uma grande variedade de sensores como ativos, passivos, vibração, nível e etc, sendo possível fazer uma completa monitoração do ambiente interno do ATM, de modo que toda a anomalia detectada dentro do ambiente protegido é enviada a uma central de monitoramento usando um celular GSM com tecnologia GPRS para transmissão de dados.



Reivindicação :

SISTEMA DE ALARME PARA TERMINAIS DE AUTOATENDIMENTO BANCÁRIO, caracterizado pelo fato de contar com setores disponíveis para a instalação de uma grande variedade de sensores como ativos, passivos, vibração, nível e etc, sendo possivel fazer uma completa monitoração do ambiente interno do ATM, de modo que toda a anomalia detectada dentro do ambiente protegido é enviada a uma central de monitoramento usando um celular GSM com tecnologia GPRS para transmissão de dados.

Anotação manual :

```
(ROOT
  (ASSUNTO_PAT SISTEMA DE ALARME PARA TERMINAIS DE AUTOATENDIMENTO BANCÁRIO),
  (CARACT_ASSUNTO caracterizado por_ o fato de contar com
    (OBJ_PAT setores disponíveis
      (CARACT_OBJ para a instalação de uma grande variedade de sensores como
       ativos, passivos, vibração, nível e etc)
    ), sendo possivel fazer uma completa monitoração de_ o ambiente interno de_ o ATM,
       de modo que toda a anomalia detectada dentro de_ o ambiente protegido é
       enviada a uma central de monitoramento usando um celular GSM com tecnologia
       GPRS para transmissão de dados
  )
)
```

Grafo extraído :