PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

PROCESSOS DE CONSTRUÇÃO AUTOMÁTICA DE TESAURO

ROGER LEITZKE GRANADA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof. Dr. Vera Lúcia Strube de Lima
Coorientadora: Prof. Dr. Renata Vieira

Porto Alegre
2011

Dados Internacionais de Catalogação na Publicação (CIP)

G748p Granada, Roger Leitzke
Processos de construção automática de tesauro / Roger Leitzke Granada. – Porto Alegre, 2011.
114 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Profª. Drª. Vera Lúcia Strube de Lima.

1. Informática. 2. Tesauros – Elaboração. 3. Indexação de Assuntos. I. Lima, Vera Lúcia Strube de. II. Título.

CDD 025.4

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada **"Processos de Construção Automática de Tesauro"**, apresentada por Roger Leitzke Granada, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 29/03/2011 pela Comissão Examinadora:

Profa. Dra. Vera Lúcia Strube de Lima – PPGCC/PUCRS
Orientadora

Prof. Dr. Marcelo Blois Ribeiro - PPGCC/PUCRS

Dra. Caroline Varaschin Gasperin – TOUCHTYPE-LTDA

Homologada em 16/11/11, conforme Ata No. 022 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

À minha família: meus pais Nilton e Gleci, e meus irmãos Rodrigo e Rossana. Dedico a vocês este trabalho por me apoiarem em todos os momentos, mesmo quando sabiam que eu estava errado. Dificilmente eu teria chegado até aqui se não fosse pela ajuda de vocês. Dedico também aos meus tios Darci e Vera que sempre me acolheram com muito boa vontade, me dando um lar quando precisei e dando alegria a minha vida quando eu estava por perto.

"A imaginação é mais importante que o conhecimento"
Albert Einstein (1879 - 1955)

"Uma paixão forte por qualquer objeto assegurará o sucesso, porque o desejo pelo objetivo mostrará os meios."
William Hazlitt (1778 – 1830)

AGRADECIMENTOS

Ao meu ex-orientador Stanley Loh, por ter acreditado no meu potencial e por ter me incentivado a trabalhar com pesquisa desde o início. Sua atitude mudou o rumo da minha vida para sempre, me fazendo descobrir o que eu realmente gosto de fazer.

À minha orientadora Vera Lúcia Strube de Lima e minha coorientadora Renata Vieira. Este trabalho deve grande parte de sua consistência ao cuidadoso trabalho de orientação e revisão realizado por vocês. Suas críticas e observações, feitas sempre de forma gentil e ponderada, me ajudaram a corrigir enfoques e informações, que, sem seu crivo teriam comprometido o resultado final do trabalho. Obrigado pelas suas conversas sempre instrutivas e que me fizeram crescer não apenas intelectualmente, mas também pessoalmente.

Aos meus amigos do mestrado, principalmente ao Igor da Silveira Wendt, que passou os dois anos de mestrado me aturando e me ajudando quando era necessário. Também aos amigos Aline Riva, Christian Quevedo, Fabiana Dorneles, Humberto Souto Junior, Michele dos Santos da Silva, Vinicius Cassol e Tiago Bortoluzzi por compartilharem ótimos momentos durante o período do mestrado, e também sempre estarem dispostos a discutir qualquer assunto.

Aos meus amigos do CPCA. Dentre eles, Henry Braun por sua paciência ao tentar me ensinar algo de útil, Fernando Castilho e Anderson Silva que sempre estiveram nas cadeiras ao lado para qualquer discussão ou brincadeira, e Alexandre Seki pelas boas conversas que tivemos.

Aos meus amigos Davide Pluda, Leonard Doiron e Sandra Paola Valencia por me mostrarem que o idioma não pode ser uma barreira para conversar com outras pessoas.

Também aos que fizeram parte deste projeto, dentre eles Mário Pereira, Thays Saldanha, Tomas Sander, Caio Northfleet e Kieran McCorry, bem como à empresa Hewlett-Packard pelo suporte financeiro e pela oportunidade de trabalhar no projeto Privacy/APAO.

Por fim e não menos importantes, aos meus amigos, que sempre estiveram comigo e de alguma forma contribuíram para a conclusão deste trabalho.

PROCESSOS DE CONSTRUÇÃO AUTOMÁTICA DE TESAURO

RESUMO

Com o progresso da tecnologia, a quantidade de informação disponível em formato digital tem aumentado rapidamente. Esse aumento se reflete na crescente importância de sistemas de Recuperação de Informações (RI) eficientes, obtendo as informações corretas quando requisitadas pelos usuários.

Tesauros podem ser associados a sistemas de RI, permitindo que o sistema realize consultas não apenas pelo termo-chave, mas também por termos relacionados, obtendo documentos relacionados, que antes não eram recuperados. A criação manual, processo longo e oneroso que dava origem aos primeiros tesauros, passa a ser realizada automaticamente, através de diferentes métodos e processos disponíveis atualmente.

Com esta motivação, este trabalho propõe estudar três processos de construção automática de tesauros. Um método utiliza técnicas estatísticas para a identificação dos melhores termos relacionados. Outro método utiliza conhecimento sintático, sendo necessário extrair, além das categorias gramaticais de cada termo, as relações que um verbo tem com seu sujeito ou objeto. O último método faz a utilização de conhecimento sintático e de conhecimento semântico dos termos, identificando relações que não são aparentes. Para isso, esse último método utiliza uma adaptação da técnica de Análise Semântica Latente.

Foram desenvolvidos estes três métodos de geração tesauros a partir de documentos do domínio de privacidade de dados. Os resultados foram aplicados a um sistema de RI, permitindo a avaliação por especialistas do domínio. Como conclusão, observamos que, em determinados casos, é melhor a aplicação de técnicas que não utilizem conhecimento semântico dos termos, obtendo melhores resultados com métodos que utilizam apenas o conhecimento sintático dos mesmos.

Palavras-chave: Tesauro, Construção de tesauro, Construção automática de tesauro.

PROCESSES FOR AUTOMATIC THESAURUS CONSTRUCTION

ABSTRACT

The advances in technology have made the amount of information available in digital format increase rapidly. This increase reflects on the importance of efficient systems to Information Retrieval (IR), getting the right information when it's requested by users.

Thesauri can be associated with IR systems, allowing the system to query not only by the key term, but also by related terms, obtaining related documents that were not retrieved. The manual construction, long and costly process that gave rise to the first thesaurus, shall be performed automatically, using different methods and processes available today.

With this motivation, this dissertation proposes to study three cases of automatic thesauri construction. One method uses statistical techniques to identify the best related terms. Another method uses syntactic knowledge, being necessary to extract, besides the grammatical categories of each term, the relations that a verb have with its subject or object. The latter method makes use of syntactic knowledge and semantic knowledge of the terms, identifying non apparent relations. For this, this latter method uses an adaptation of the Latent Semantic Analysis technique.

We developed three methods for automatic thesaurus construction using documents from the field of data privacy. The results were applied to an IR system, allowing the evaluation by domain experts. In conclusion, we observed that, in certain cases, it's better to apply techniques that do not use semantic knowledge of the terms, obtaining better results with methods that use only the syntactic knowledge of them.

Keywords: Thesaurus, Thesaurus construction, Automatic thesaurus construction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CPCA	Centro de Pesquisa em Computação Aplicada
HP	*Hewlett-Packard*
IM	Informação Mútua
LSA	*Latent Semantic Analysis* (Análise Semântica Latente)
LSI	*Latent Semantic Indexing* (Indexação Semântica Latente)
NSP	*Ngram Statistical Package* (Pacote estatístico Ngram)
PF-IBF	*Path Frequency - Inversed Backward link Frequency* (Frequência do caminho – Frequência inversa para o *link* anterior)
POS	*Part Of Speech* (Categoria gramatical)
PUCRS	Pontifícia Universidade Católica do Rio Grande do Sul
RI	Recuperação de Informações
SVD	*Singular Value Decomposition* (Decomposição em Valores Singulares)
TF-IDF	*Term Frequency – Inverse Document Frequency* (Frequência do termo – Frequência Inversa do Documento)
TOEFL	*Test Of English as a Second Language* (Teste de Inglês como Segunda Língua)
TREC	*Text REtrieval Conference* (Conferência em Recuperação de Textos)
UNESCO	*United Nations Educational, Scientific and Cultural Organization* (Organização das Nações Unidas para a Educação, Ciência e Cultura)

SUMÁRIO

1. INTRODUÇÃO

Com o avanço da tecnologia, a quantidade de informação disponível em formato eletrônico tem aumentado significativamente. Com isso, a área de Recuperação de Informações (RI) baseada em conteúdo tem se tornado uma importante área de pesquisa. Através de técnicas variadas, a área de RI busca encontrar o conteúdo relevante que responda a uma consulta de um usuário, por exemplo.

Devido à existência da sinonímia entre termos e devido à habilidade das pessoas para descrever a mesma ideia de diferentes maneiras, a RI torna-se pouco eficiente caso uma busca seja realizada utilizando apenas o termo procurado. Para maior eficácia na RI, pode ser feita a associação de termos semelhantes, ou seja, termos semanticamente relacionados ao termo-chave que está sendo procurado.

Um tesauro é um conjunto de termos que são associados a termos-chave de acordo com sua similaridade semântica. Pelo fato de um tesauro se basear na identificação de termos semanticamente similares a um termo-chave, ele pode ser utilizado na RI, associando aos termos procurados outros termos semanticamente similares, e recuperando não só os documentos que contêm o termo específico, mas também outros documentos relacionados.

Inicialmente a construção de tesauros era feita manualmente. Porém a construção manual de tesauros exige muito tempo e este recurso é de difícil manutenção. Por outro lado, o aumento da disponibilidade de textos em formato digital pode permitir a criação automática de tesauros a partir de uma coleção de documentos.

Métodos para a criação automática de tesauros vêm há muito tempo sendo estudados, alguns utilizando apenas técnicas estatísticas, outros utilizando conhecimento morfológico e sintático, outros utilizando, além de conhecimento morfológico e sintático, também conhecimento semântico para a identificação dos termos semelhantes.

Métodos mais recentes têm utilizado o grande volume de dados textuais disponíveis na internet (como a Wikipédia[1]) para a criação de tesauros e outros métodos, ainda, utilizam o aprendizado de máquina para a melhor identificação dos termos semelhantes.

[1] *http://www.wikipedia.org/*

Embora diversos métodos tenham sido propostos, ainda é uma tarefa difícil fazer a geração automática de um tesauro de forma que se encontrem os termos que melhor cubram o escopo da coleção.

Esta dissertação de mestrado propõe uma análise comparativa entre técnicas de construção automática de tesauros a partir de textos. Esta é a motivação principal deste trabalho, cujo contexto e escopo são apresentados nas próximas seções.

1.1 Contextualização

Atualmente empresas têm demonstrado grande interesse na criação de sistemas para o gerenciamento de políticas de privacidade. Estes sistemas ajudam o projetista a descobrir os devidos cuidados que devem ser tomados em relação à privacidade de dados, quando novos projetos são criados.

Para ajudar um projetista os sistemas utilizam uma base de conhecimento contendo leis, normas, regulamentos etc. relacionados à privacidade de dados que são consultados, conforme as descrições de um determinado projeto.

Embora estas bases de conhecimento sejam criadas manualmente por especialistas, elas têm difícil manutenção quando alguma lei, norma ou regulamento é modificado. Para identificar quais partes da base de conhecimento são afetadas quando ocorre uma destas mudanças, um sistema de RI pode ser disponibilizado, recuperando documentos de acordo com os termos-chave especificados.

Dessa forma, caso uma lei que descreve, por exemplo, como serão tratadas informações que podem identificar uma pessoa (do inglês, “*Personal Identifiable Information*”) seja modificada, o sistema de RI é capaz de recuperar todos os documentos em que este termo aparece.

Embora o sistema recupere documentos contendo o termo específico, ele não recupera todos os documentos a respeito das informações que podem identificar uma pessoa, pois em muitos documentos esse tema é tratado por outros termos, como, por exemplo, “*personal information*”, “*personal data*” ou até mesmo o acrônimo “*PII*”.

Para fazer a identificação dos termos similares a um termo-chave, deve ser introduzido um tesauro no sistema de RI, de forma que, quando o sistema recuperar documentos, ele obtenha não só documentos que tenham o termo-chave associado,

como também documentos que tenham associados os termos semanticamente relacionados.

O projeto de pesquisa onde este trabalho de mestrado se insere, é denominado Privacy/APAO, e conduzido através de uma parceria PUCRS/HP no período de Março de 2009 à Dezembro de 2010.

1.2 Definição do projeto de pesquisa

Nossa questão de pesquisa se refere aos processos de criação automática de tesauro, procedendo a uma análise e avaliação de métodos encontrados na literatura. Após revisão bibliográfica minuciosa, foram escolhidos os métodos de Kaji *et al*. [KMAY00], Grefenstette [Gre94] e Yang e Powers [YP08].

Assim, apresentamos e comparamos três métodos de construção, sendo um método baseado em técnicas estatísticas, um método baseado em análise de informação morfológica e sintática do corpus e um método que utiliza, além informação morfológica e sintática do corpus, a aplicação de uma adaptação da técnica de Análise Semântica Latente (do inglês, *Latent Semantic Analysis* - LSA).

Métodos mais recentes, como os descritos no trabalho de Li *et al*. [LSP09] e no trabalho de Xu e Yu [XY10], tendem a fazer a geração automática do tesauro utilizando o aprendizado de máquina, porém esse tipo de técnica não será abordado no presente trabalho por não dispormos de um corpus de treino.

Os resultados de cada um dos três métodos de geração automática de tesauros foram inseridos em um sistema de RI, permitindo assim uma avaliação dos resultados, que é discutida ao final deste volume.

1.3 Organização do trabalho

O restante do texto da dissertação está organizado da seguinte forma. O Capítulo 2 apresenta a fundamentação teórica deste trabalho, descrevendo conceitos e definições referentes ao modo como são realizadas as construções de tesauros e suas avaliações, bem como uma breve explicação sobre conceitos de Análise Semântica Latente, que é aplicada em um dos métodos de construção automática de tesauros.

Trabalhos relacionados à construção de tesauros por métodos estatísticos, sintáticos e com a utilização de LSA são apresentados no Capítulo 3. Os recursos

utilizados, ferramentas utilizadas e a aplicação desenvolvida para a construção de tesauros são apresentados no Capítulo 4. O processo de escolha dos termos para avaliação, a aplicação da ferramenta em um sistema de RI, e o processo de avaliação podem ser vistos no Capítulo 5. No Capítulo 6 são apresentados e discutidos os resultados obtidos pela avaliação com especialistas de domínio.

Finalmente, o documento é encerrado no Capítulo 7, onde são apresentadas as considerações do trabalho e introduzidas sugestões de trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Tesauros

2.1.1 Breve histórico

Segundo Knapp em [Kna00], o termo "tesauro" tem origem da palavra latina *thesaurus* e da palavra grega *thesaurós*, e significando tesouro ou depósito. A partir do século XVIII passou a ter o significado de "depósito de conhecimento" como um dicionário ou enciclopédia. O termo tesauro tem sido usado durante muitos séculos para designar léxico ou tesouro de palavras. Por tesouro de palavras não se entende apenas quantidade de palavras, mas sim riqueza de conceitos e relações semânticas que devem existir entre elas.

Uma das primeiras obras a incluir o termo no seu título foi *Thesaurus linguae Romanae et Britannicae* de autoria de Thomas Cooper [Coo69], publicada em 1565. Porém a palavra popularizou-se a partir da publicação de Peter Mark Roget chamada *Thesaurus of English Words and Phrases* em 1852. Essa publicação tem o objetivo de classificar e organizar os termos de forma a facilitar a expressão das ideias [RS68]. Assim, essa publicação passa a organizar as palavras não mais em ordem alfabética, mas sim pelas ideias que elas expressam.

Segundo Campos e Gomes [CG06], em 1950, Hans Peter Luhn, do centro de pesquisa da IBM nos Estados Unidos, percebeu que para encontrar a palavra/ideia mais adequada a uma recuperação de informação, não adiantava apenas ter uma lista de palavras em ordem alfabética. Era necessário utilizar uma estrutura mais complexa, onde as palavras deveriam ter referências cruzadas de forma a evidenciar quais ideias estavam interconectadas. Deu o nome de *thesaurus* a esta nova lista, influenciado pelo trabalho de Roget, que, em seu dicionário analógico, o define da seguinte forma: “a revisão de um catálogo de palavras de significado análogo vai sugerir, com frequência, por associação, outras sucessões de pensamentos”.

Ainda segundo Campos e Gomes [CG06], isso fez com que fosse criado um novo tipo de linguagem documentária, conhecido como tesauro de recuperação de informação, que veio para auxiliar sistemas que utilizavam um único termo (unitermo). Outras listas de termos que continham algumas relações entre eles também passaram a se chamar tesauros.

Com o advento da internet, os tesauros têm sido utilizados para recuperar documentos que o usuário necessita. Pois, assim como o tesauro ajuda um bibliotecário a encontrar documentos através de palavras-chave, ele também auxilia um usuário de sistemas de informação a encontrar os documentos que necessita.

2.1.2 Alguns conceitos para tesauros

Uma das primeiras conceituações de tesauro foi dada pela UNESCO (*United Nations Educational, Scientific and Cultural Organization*) [Une70] na década de 70 e caracteriza tesauro para a área de ciência da informação, organizando essa caracterização em dois aspectos de forma que pudesse atender tanto a área de elaboração de tesauro quanto à área de recuperação de informação.

Um desses aspectos é expresso de acordo com a estrutura do tesauro: descreve o tesauro como um vocabulário controlado e dinâmico de termos relacionados semântica e genericamente, cobrindo um domínio específico do conhecimento. O outro aspecto se volta a um tesauro quanto a sua função, definindo-o como um dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários, numa linguagem de sistema mais restrita.

Outras conceituações também têm sido dadas ao termo tesauro. Wilks *et al*. [WSG96] caracterizam tesauro como a organização de listas de palavras parcialmente sinônimas que explicam ou definem o significado de uma palavra somente pelos seus sinônimos, sem a necessidade de explicar o seu significado. Um exemplo desse entendimento é a utilização dos *synsets* da Wordnet[2] para a descoberta do significado de um termo.

Para Aitchison [AGB02] tesauro é um vocabulário controlado, formalmente organizado de forma que, a priori, os relacionamentos entre conceitos são feitos explicitamente. Aitchison ainda enfatiza que um tesauro tem duas propostas, sendo a primeira delas a recuperação de informações, e a segunda, auxiliar o entendimento geral de uma determinada área, fornecendo “mapas semânticos”, mostrando os inter-relacionamentos entre os conceitos e ajudando a encontrar definições para os termos.

---

[2] *http://wordnetweb.princeton.edu/perl/webwn*

Outras conceituações como a dada por Kilgarriff e Yallop [KY00] são mais simplistas e caracterizam tesauros apenas como um recurso no qual as palavras com significados semelhantes são agrupadas. Observa-se que muitas definições têm sido dadas ao termo tesauro, porém a maioria delas dá ênfase às relações entre as palavras (seja com similaridade semântica ou sinonímica) e a descoberta do significado de uma palavra através das palavras a ela relacionadas.

Uma explicação para o entendimento de um tesauro, ou como ele poderia ser construído, pode ser visto no trabalho de [Lin98], que diz:

> *A bottle of tezgüino is on the table.*
> *Everyone likes tezgüino.*
> *Tezgüino makes you drunk.*
> *We make tezgüino out of corn.*

A partir do exemplo apresentado, pode se entender que *tezgüino* sugere algum tipo de bebida alcoólica feita de milho. Dessa forma, poderíamos associar a esse termo, palavras como cerveja, vinho, cachaça etc., formando dessa maneira o tesauro.

2.1.3 Aplicações

Explicados os conceitos e obtido o entendimento do que é um tesauro, pode-se observar que ele pode ser utilizado em diversas aplicações, tendo sido inicialmente utilizado em sistemas de recuperação de informações. O trabalho de Jing e Croft [JC94] faz a criação automática de um tesauro que é utilizado em um sistema chamado PhraseFinder. O tesauro é utilizado para fazer a expansão de consultas, isto é, associar termos relacionados aos termos de uma consulta em um sistema de recuperação de documentos.

Um trabalho semelhante, porém em para a língua portuguesa, é apresentado por Pizzato e de Lima [PL03], que utiliza um tesauro construído automaticamente para fazer a expansão de consultas na recuperação de documentos.

Além de RI, os tesauros também têm sido utilizados para outras finalidades. O trabalho de Heilman e Eskenazi [HE07], por exemplo, faz a aplicação de um tesauro gerado automaticamente a partir de um corpus, em um sistema de perguntas chamado REAP. O sistema REAP provê aos estudantes um ambiente de aprendizagem apresentando um conjunto de textos e exercícios que visam o aprimoramento do conhecimento de que o estudante dispõe na língua inglesa.

O objetivo do sistema é permitir aos alunos estudarem as palavras desconhecidas que aparecem em um texto, ao invés de apenas listar o seu significado. Para isso, o sistema faz a geração de perguntas e itens para resposta, conforme apresentado no Fragmento 1, onde os termos em itálico são a resposta correta para a pergunta.

Fragmento 1. Exemplo de pergunta e resposta do sistema REAP (adaptado de [HE07]).

Which set of words are most related in meaning to "reject"?

A. pray, forget, remember
B. invest, total, owe
C. *accept, oppose, approve*
D. persuade, convince, anger

Nesse tipo de sistema o tesauro é utilizado para fazer a obtenção dos termos relacionados (no caso do Fragmento 1, termos opostos), criando a partir deles os itens de resposta. A utilização de um tesauro criado automaticamente a partir de um corpus se deu devido a existência de termos raros, que não conseguiam ser cobertos por tesauros gerais como o WordNet.

Outra aplicação dos tesauros gerados automaticamente é a categorização de documentos, como apresentado no trabalho de Wang *et al*. [WHZC09], que faz a geração de um tesauro baseado na Wikipédia. Com o tesauro construído, documentos são submetidos ao sistema de classificação de textos.

São procurados os conceitos da Wikipédia dentro do documento a ser classificado e, para cada um dos conceitos encontrados, são associados sinônimos, hipônimos e termos relacionados encontrados no tesauro, dentro do documento, enriquecendo assim a representação do documento. Dessa forma, documentos relacionados que originalmente não compartilhavam termos em comum são enriquecidos com os mesmos conceitos e classificados, sendo o resultado dessa nova classificação mais próximo da correta categorização dos documentos.

Para a língua portuguesa, o trabalho de Moraes e Lima [ML08] faz a extração de termos a partir de documentos da seção de esportes do jornal Folha de São Paulo. Para a extração dos termos é utilizada a técnica de Frequência do Termo – Frequência Inversa do Documento (do inglês, *Term Frequency – Inverse Document Frequency* – TF-IDF) para termos simples, e a técnica de *C-Value* [FAT98] para documentos compostos. Após a extração dos termos os autores fazem a clusterização dos mesmos, agrupando termos mais similares.

Por fim, trabalhos como o de Li *et al*. [LSP09], que faz a utilização de um tesauro construído automaticamente para a categorização de documentos baseado em redes neurais; e de Xu e Yu [XY10], que faz a criação automática de um tesauro para a filtragem de *spams*, reforçam a ideia de quão importante é a criação de tesauros.

2.1.4 Construção de tesauros

Existem diferentes formas de construção de tesauros, entre os métodos de apresentados a seguir, alguns utilizam métodos estatísticos, outros conhecimento sintático dos termos, outros ainda utilizam, além da combinação de ambos, um conhecimento semântico dos termos. Alguns tesauros podem ser criados de forma manual e outros de forma automática. Algumas técnicas podem ser genéricas, isto é, sem um domínio específico, e outras orientadas a domínio, utilizando assim um corpus específico para a construção. A seguir são apresentadas algumas dessas formas de construir tesauros.

2.1.4.1 Construção manual de tesauros

Inicialmente os tesauros eram criados de forma manual, porém esse tipo de construção demanda um conhecimento humano sobre o assunto na escolha dos melhores termos para compor cada conceito. Por exigir uma demanda do ser humano esse tipo de construção necessita muito tempo e esforço. Esses tipos de tesauros também são de difícil manutenção, pois a língua é dinâmica e está em constantes mudanças, isto é, novas terminologias são adicionadas à língua, palavras do domínio técnico passam a ser de domínio comum, e outras ainda tornam-se obsoletas ou temporariamente impopulares.

Jing [JC94] coloca que existem dois tipos de tesauros construídos manualmente. O primeiro deles é construído para um propósito geral e baseado em termos, como o tesauro de Roget e a Wordnet, e contém termos relacionados como antônimos, sinônimos etc.

O outro tipo é orientado a recuperação de informações como, por exemplo, o INSPEC, LCSH (*Library of Congress Subject Headings*) e MeSH (Medical Subject Headings). O último contém relações do tipo BT (“termos mais genéricos”, do inglês *Broader Term*), NT (“termos mais específicos”, do inglês *Narrower Term*), UF (“termos

usados para", do inglês *Used For*) e RT ("termos relacionados", do inglês *Related To*) e podem ser genéricos ou específicos, dependendo das necessidades de sua construção.

Para a criação manual de tesauros foram propostas padronizações, possibilitando assim a ampliação da utilidade de um tesauro, pois ao invés de estar limitado somente a uma aplicação, ele pode ser utilizado em várias aplicações, inclusive aplicações interligadas. O trabalho de Aitchison *et al*. [AGB02] descreve em detalhes como deve ser feita a construção manual de tesauros, bem como as normas para tal.

Alguns dos tesauros construídos manualmente e mais utilizados atualmente são descritos abaixo:

- Tesauro de Roget

Apesar do tesauro de Roget [RS68] ter sido desenvolvido em 1852 e finalizado 50 anos depois, ele ainda é amplamente utilizado como veremos na subseção 2.1.5 que trata de avaliação de tesauros. Esse tesauro agrupa os itens lexicais em seis grandes classes: (i) *abstract relations*, (ii) *space*, (iii) *matter*, (iv) *intelect*, (v) *volition* e (vi) *affections*. Essas classes são divididas em seções e cada seção é subdividida, resultando em um total de 1000 grupos semânticos.

O tesauro de Roget se diferencia dos dicionários, pois ele é organizado de acordo com o significado dos termos e não segundo a ordem alfabética. Esse tesauro foi construído com o objetivo de ajudar o usuário a encontrar palavras que representem uma ideia desejada. Essas ideias são expressas através de palavras relacionadas, normalmente sinônimos, mas também podem conter antônimos ou termos relacionados.

- WordNet

Outro tipo de tesauro construído manualmente e muito utilizado na literatura é a WordNet [Fel98]. A WordNet é uma base de dados lexical para a língua inglesa, desenvolvida sob a direção de George A. Miller na Princeton University. Nela substantivos, verbos, adjetivos e advérbios são agrupados em um conjunto cognitivo de sinônimos (*synsets*), onde cada um expressa um conceito distinto.

A WordNet é mais parecida com um tesauro do que com um dicionário devido à maneira como organiza os *synsets*. Os *synsets* são interconectados através de relações semântico-conceituais e lexicais. Essas relações variam de acordo com o tipo de palavra como mostrado na Tabela 2.1.

Segundo Miller em [Mil95], essas relações foram escolhidas para compor a WordNet pois elas são amplamente utilizadas na língua inglesa e com isso são bastante familiares à maioria das pessoas, não sendo necessário um conhecimento avançado em linguística para entendê-las.

Tabela 2.1. Relação entre termos na WordNet (adaptado de [Fel98])

Substantivos	
Sinônimos	$termo_x$ é um sinônimo do $termo_y$ se o $termo_x$ e o $termo_y$ compartilham de um mesmo sentido (automóvel é sinônimo de carro);
Antônimo	$termo_x$ é um antônimo do $termo_y$ se o $termo_x$ e o $termo_y$ compartilham de sentidos opostos (certo é antônimo de errado);
Hiperônimos	$termo_x$ é um hiperônimo do $termo_y$ se o $termo_y$ *é um tipo de* $termo_x$ (veículo é hiperônimo de carro);
Hipônimo	$termo_y$ é um hipônimo do $termo_x$ se o $termo_y$ *é um tipo de* $termo_y$ (carro é hipônimo de veiculo);
Holônimo	$termo_y$ é um holônimo do $termo_x$ se o $termo_x$ *é parte de* $termo_y$ (casa é holônimo de janela);
Merônimo	$termo_x$ é um merônimo do $termo_y$ se o $termo_x$ *é parte de* $termo_y$ (janela é merônimo de casa);
Termos coordenados	Aqueles que compartilham o mesmo hiperônimo (ônibus e carro são termos coordenados, pois têm o mesmo hiperônimo, ônibus);
Verbos	
Hiperônimo	O $verbo_y$ é um hiperônimo do $verbo_x$ se a ação $verbo_x$ *é um tipo de* $verbo_y$ (perceber é um hiperônimo de ouvir);
Tropônimo	O $verbo_y$ é um tropônimo do $verbo_x$ se a ação $verbo_x$ *é um modo particular de* $verbo_y$ de alguma maneira (sussurrar é um tropônimo de falar);
Implicação	O $verbo_y$ *é implicado pelo* $verbo_x$ se para fazer o $verbo_x$ eu devo ter feito primeiro o $verbo_y$ (roncar é implicado por dormir);
Termos coordenados	Aqueles termos que compartilham um mesmo hiperônimo (sussurrar e gritar são termos coordenados, pois têm o mesmo hiperônimo, falar);
Adjetivos e advérbios	
Antônimo	Ex. forte é antônimo de fraco;
Sinônimo	Ex. triste é sinônimo de infeliz.

A Wordnet é amplamente utilizada, não apenas como consulta para a obtenção dos termos relacionados a um determinado termo, mas também como tesauro de referência

para avaliação de tesauros, conforme será apresentado na subseção 2.1.5 referente à avaliação de tesauros.

2.1.4.2 Construção automática de tesauros

Em Recuperação de Informação (RI) os tesauros têm um papel importante e, tratando de um domínio específico, eles auxiliam ainda mais os sistemas de RI, pois podem fornecer desde um controle sobre o vocabulário na indexação dos termos, até adicionar termos semelhantes que visem melhorar a RI. Por este interesse em auxiliar os sistemas como os de RI, muitas vezes as pessoas se veem confrontadas com a questão de criar e manter automaticamente um tesauro de um domínio específico.

Um tesauro pode ser construído automaticamente sem a necessidade de um corpus (isso ocorre, por exemplo, nos casos em que um tesauro é apenas traduzido), ou através de um corpus.

A construção automática de um tesauro se baseia na identificação, de forma automatizada, dos relacionamentos semânticos entre as palavras, encontrando assim palavras mais similares a uma palavra-chave. Essa identificação automática pode se dar sem o uso de um corpus (como no caso de tesauros construídos apenas pela tradução de outros tesauros), ou com o uso de corpus.

Quando o processo é de construção automática, Ito *et al*. [INHN08] classificam os tesauros em dois tipos: tesauros relacionais e tesauros associativos. A seguir, desenvolvemos esses dois conceitos.

- Tesauros relacionais

Segundo Ito *et al*. [INHN08], são tesauros que definem explicitamente os relacionamentos (como "é-um" e "é-parte-de"). Kaji *et al*. [KMAY00] generalizam essa classificação, denominando esses tipos de tesauros de taxonômicos. Dessa forma, tesauros relacionais (ou taxonômicos) são aqueles que apresentam uma estrutura taxonômica entre seus termos, podendo ser encontradas nesta estrutura relações como meronímia, hiponímia, hiperonímia etc..

Para a geração automática de tesauros taxonômicos, diversas abordagens vêm sendo adotadas. No trabalho de Hearst [Hea92] é feita a extração de hipônimos de um corpus através da busca por padrões. Um desses padrões, por exemplo, é encontrar a

expressão “*such as*” entre dois sintagmas nominais (*SN_1 such as SN_2*), o que pode indicar que o sintagma nominal SN_2 é um hipônimo de SN_1.

Outros trabalhos que exploram a geração de tesauros taxonômicos são descritos por Ponzetto e Strube [PS07], Sumida e Torisawa [ST08] e Wang *et al*. [WHZC09]. Esses trabalhos fazem a extração de termos e construção de uma taxonomia (com relações “*is a*”, “*not is a*”, por exemplo), baseado na estrutura de artigos da Wikipédia.

- Tesauros associativos

Segundo Kaji *et al*. [KMAY00], tesauros associativos utilizam a associação semântica entre termos na sua construção. Assim, pode-se dizer que o significado de um termo é dado pelos outros termos que o modificam. Esse tipo de tesauro pode ser visto como um grafo, onde os conceitos são representados pelos nodos e as relações entre os conceitos são representadas pelas arestas.

No ano de 1954, Zelig S. Harris publicou em seu artigo "*Distributional Structure*" [Har54] a hipótese de que palavras tendem a ter o mesmo significado se compartilham de contextos semelhantes. Baseados nessa hipótese, diversos trabalhos vêm fazendo a identificação dos termos que compartilham os mesmos contextos, para gerar um tesauro.

Alguns, como o de Kaji *et al*. [KMAY00], utilizam o valor de Informação Mútua entre os termos para a obtenção do tesauro, enquanto que outros trabalhos como o de Grefenstette [Gre94] passam a utilizar uma métrica de similaridade entre contextos sintáticos, para a identificação dos termos. Ainda, trabalhos como o de Yang e Powers [YP08] vão além do trabalho de Grefenstette e utilizam uma adaptação da técnica de Análise Semântica Latente para obter valores relacionados aos termos antes de aplicar uma métrica de similaridade entre os mesmos.

O trabalho de Kilgarriff e Yallop [KY00] faz a descrição de um algoritmo simplista para a construção automática de tesauros associativos, que é apresentado no Fragmento 2.

Fragmento 2. Algoritmo para a construção automática de tesauros (adaptado de [KY00])

```
For each content word in the corpus
    for each other content word,
        find how often both occur within k words (or characters) of each other.
```

Dessa forma, se o corpus é composto de *n* termos, cada termo é representado por um vetor de tamanho *n*. A similaridade entre dois vetores pode ser computada através de

uma fórmula de similaridade, identificando assim os termos mais similares para cada um dos termos do corpus.

Além desses trabalhos, diversos outros vêm sendo propostos na tentativa de identificação dos melhores relacionamentos entre as palavras [Cro88, CH90, Gre94, Lin98, KMAY00, Gas01, GL03, NHN07, CC07, AMS08, Bin08, INHN08, KHT08, YP08. LSP09, XU10]. Porém, mesmo sendo a criação automática de tesauros um assunto antigo, cabe ressaltar que ainda é um desafio encontrar os melhores relacionamentos entre as palavras de forma que o tesauro contenha termos que melhor cubram o escopo dos documentos da coleção.

2.1.5 Avaliação de tesauros

Rapp [Rap04] coloca que, ao computar a similaridade semântica entre termos, é desejável que seja feita uma avaliação dos resultados obtidos e, assim, diferentes métodos de avaliação de tesauros construídos automaticamente têm surgido.

Segundo Yang e Powers [YP08] não é uma tarefa trivial fazer uma avaliação de tesauro com a ausência de um *benchmark*. Pode ser feita uma avaliação subjetiva, observando a similaridade distribucional e avaliando assim a qualidade dos grupos de termos formados. Porém uma avaliação assim é muito custosa, pois necessita de *experts* para a correta medição e se torna inviável para um corpus muito grande.

Outra possibilidade de avaliação de um tesauro é através da comparação dos termos resultantes com outros tesauros ou extratos de tesauros já existentes. Esses outros tesauros ou fontes lexicais já existentes são um *gold standard*, ou tesauro de referência. Grefenstette [Gre94], por exemplo, usou o tesauro de Roget (versão de 1911), o tesauro de Macquarie e o Webster’s 7th Dictionary como estruturas de referência.

Nessa avaliação o autor verificava se dois termos estavam localizados no mesmo tópico nos tesauros de Macquarie e de Roget, ou compartilhando duas ou mais definições no Webster’s 7th Dictionary. Nesse caso ele era contado como um termo válido, sendo assim um sinônimo ou termo semanticamente relacionado.

Yang e Powers [YP08] fazem a avaliação de modo semelhante, porém é utilizado como *gold standard*, além do tesauro de Roget, a WordNet. Assim, além das relações de sinônimos/antônimos providos pela WordNet, conseguem cobrir termos que não têm um

relacionamento definido, mas que estão associados a um determinado tópico do tesauro de Roget.

Outros tesauros utilizados como *gold standard* são o Moby Thesaurus[3], utilizado no trabalho de Heilman e Eskenazi [HE07] e o conjunto de sinônimos do *Test of English as a Foreign Language* (TOEFL), utilizado nos trabalhos de Rapp [Rap04, Rap08]. Uma consideração a ser feita quanto a esse tipo de técnica de avaliação é que, em domínios muito específicos, pode ocorrer inviabilidade de utilizá-la.

2.2 Análise Semântica Latente (LSA)

Nesta seção apresentaremos uma técnica que visa fazer um agrupamento semântico entre termos que compartilham os mesmos documentos. Essa técnica é chamada de Análise Semântica Latente pois visa encontrar relações entre termos que não são aparentes quando observados somente do ponto de vista de suas frequências nos documentos.

Essa técnica é apresentada devido ao trabalho de Yang e Powers [YP08] utilizar uma adaptação da mesma para fazer a construção automática de um tesauro. A seguir são descritas algumas definições de LSA e uma breve descrição do seu funcionamento.

2.2.1 Definições

A Análise Semânica Latente (do inglês, *Latent Semantic Analysis* – LSA) é uma técnica que surgiu motivada pela deficiência do modelo vetorial para lidar com sinônimos. Sahlgren [Sah06] coloca que em um sistema de RI que utiliza um modelo vetorial dificilmente conseguiremos recuperar documentos com a palavra “*ship*” se utilizarmos como termo para a pesquisa a palavra “*boat*”.

Por fazer um agrupamento dos termos através das relações semânticas existentes entre os mesmos, a técnica de LSA pode ser utilizada para associar termos semelhantes ao termo que está sendo consultado. Dessa forma ao invés de fazer a recuperação de documentos que contenham apenas o termo “*ship*”, o sistema poderá recuperar documentos que contenham o termo “*boat*”, se o mesmo esteja relacionado semanticamente com o termo “*ship*”.

---

[3] *http://icon.shef.ac.uk/Moby/mthes.html*

Inicialmente essa técnica foi denominada de LSI, referindo-se a Indexação Semântica Latente (do inglês, *Latent Semantic Indexing*) e foi apresentada no trabalho de Dumais *et al*. [DFL+88], porém os autores passaram a referenciá-la também por LSA.

Manning e Schütze [MS99] colocam que a LSA é uma técnica que projeta termos e documentos dentro de um espaço com “dimensões de semântica latente”, isto é, projeta os termos que coocorrem em uma mesma dimensão, e os que não coocorrem com ele, em diferentes dimensões. Com isso, a LSA é uma técnica que faz uma redução de dimensões, partindo de várias dimensões, que representam termos e documentos, e condensando-as em poucas dimensões “latentes” que colapsam termos e documentos com vetores de contexto similares.

Berry *et al*. [BDO95] colocam que o LSA assume que existe uma estrutura latente, ou não aparente, no uso das palavras, e que essa estrutura é parcialmente escondida pela variabilidade da escolha das palavras na construção de um texto.

2.2.2. LSA na construção automática de tesauros

Segundo Landauer *et al*. [LFL98], os passos para a descoberta do significado das palavras são:

1) Fazer uma representação do corpus em uma matriz, onde cada linha representa uma única palavra do corpus e cada coluna representa uma passagem ou contexto em que as palavras ocorrem. Cada célula contém a frequência de ocorrência da palavra (linha) em determinado contexto (coluna).

Neste passo ainda pode ser adicionado um peso que expressa a importância do termo na passagem, e também a importância que o termo tem no domínio em geral. Esse peso é calculado baseado em entropias das coocorrências como pode ser visto no trabalho de Dumais [Dum93].

2) Após, deve-se aplicar sobre a matriz a técnica de Decomposição em Valores Singulares (do inglês, *Singular Value Decomposition* - SVD) [GK65]. A SVD é um modelo matemático que pode ser visto como uma análise de coocorrência de termos através de um método de redução de dimensões de uma matriz. Através desse modelo, tenta-se encontrar relação entre palavras que podem não ocorrer em uma mesma passagem. Assim, mesmo se uma palavra pal_x não aparece em um documento com a palavra pal_y,

ela pode ter uma correlação alta com pal_y, desde que elas compartilhem de contextos com significados similares.

Conforme apresentado por Sahlgren [Sah06], um argumento para a utilização da SVD é que o espaço resultante não tem apenas as relações encontradas “na superfície”, isto é, aquelas que estão na matriz de termos x documentos, mas também são encontradas relações “latentes”, isto é, relações que estão implícitas na matriz. Isso ocorre pois com a aplicação da técnica de SVD, termos com padrões de coocorrência similares passam a ser agrupados juntos.

Manning e Schutze [MS99] explicam que para encontrar os contextos similares, parte-se de uma matriz original A_{txd} com t linhas e d colunas, onde t representa os termos dos documentos e d representa os documentos do corpus. Com isso, o conteúdo do elemento a_{ij} será a frequência do termo i no documento j.

O SVD decompõe a matriz A_{txd} e a representa através de três matrizes U_{txn}, Σ_{nxn} e V_{dxn}, onde $n = min(t,d)$. A matriz V é rotacionada sobre sua diagonal principal ($V_{ij} = V^T_{ji}$) obtendo-se V^T_{nxd}, com isso obtem-se:

$$A_{txd} = U_{txn}\ \Sigma_{nxn}\ e\ V^T_{nxd} \quad (1)$$

As matrizes U e V^T contêm colunas ortonormais, isto é, os vetores-coluna são ortogonais entre eles. Assim, se uma matriz C tem colunas ortonormais, então $C^TC = I$, onde I é uma matriz diagonal com valores singulares iguais a 1 na diagonal principal e valores 0 no resto. Resumindo, da equação 1, obtém-se:

- A: Matriz original de dimensões $m \times n$ que relaciona os termos com os documentos;
- U: Matriz ortogonal de dimensões $m \times k$, onde k é um valor entre m e n;
- Σ: Matriz com valores singulares de dimensões $k \times k$, onde os valores singulares estão ordenados de forma decrescente, sendo nulos nas últimas linhas da matriz;
- V^T: Matriz ortogonal de dimensões $k \times n$, onde o T indica que é a matriz V transposta (isto é, as colunas são linhas e as linhas viram colunas).

Um problema ao se trabalhar com matrizes termos x documentos é que 99% das células conteriam valor zero, já que a grande maioria das palavras somente ocorrem em um número muito limitado de contextos. Porém, com a aplicação da SVD a matriz resultante será uma matriz de dimensões reduzidas, se comparada à matriz original.

Essa decomposição pode obter como resultado três tipos de matrizes:

Matriz de comparação termo-termo:

$$\hat{A}_k = U \Sigma^2 U^T \quad (2)$$

Matriz de comparação termo-documento:

$$\hat{A}_k = U \Sigma V^T \quad (3)$$

Matriz de comparação documento-documento:

$$\hat{A}_k = V \Sigma^2 V^T \quad (4)$$

3) Por fim, uma medida de similaridade é utilizada no espaço dimensional reduzido, completando assim o processo de inferência das palavras ou passagens. Normalmente neste processo é utilizada a medida da distância entre os dois vetores (utilizando a abordagem do Cosseno).

3. TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos que estão relacionados à construção automática de tesauros a partir de corpus. Para facilitar o entendimento das técnicas utilizadas, os trabalhos foram separados em três grupos: trabalhos que utilizam técnicas baseadas em janelas, trabalhos que utilizam técnicas baseadas em sintaxe, e por fim, trabalhos que, além de necessitarem da anotação sintática do corpus, ainda procuram por relações semânticas entre os termos para a identificação dos termos relacionados na construção do tesauro.

3.1 Técnicas baseadas em janelas

As técnicas aqui descritas, embora utilizem em sua grande parte apenas métodos estatísticos para fazer a criação de um tesauro, incluem aquelas em que é utilizado conhecimento linguístico básico, onde os termos do corpus podem ser anotados com suas categorias gramaticais por um etiquetador, também conhecido como *Part of Speech (POS) tagger*. Porém, não foram incluídos trabalhos onde um conhecimento mais profundo da língua é necessário, exigindo a análise sintática das frases do corpus.

Firth afirmou em seu trabalho "*A synopsis of linguistic theory 1930-1955*" ([Fir75] apud [CH90]) que você deve conhecer uma palavra por aquelas que a acompanham. A ideia dos trabalhos que utilizam técnicas baseadas em janelas é justamente essa, conhecer as palavras através das palavras que compartilham os mesmos contextos. Nas subseções seguintes são apresentadas breves descrições de alguns trabalhos que utilizam tais técnicas.

3.1.1. Crouch [Cro88]

O trabalho de Crouch [Cro88], seguido por Crouch e Yang [CY92], apresenta a criação automática de um tesauro global. Por tesauro global entende-se um tesauro que é criado uma vez e pode ser usado num ambiente de recuperação de informação para indexar tanto documentos como consultas. Um tesauro local, por sua vez, é gerado dinamicamente a cada consulta que é realizada no sistema.

Nesse trabalho, a criação automática de um tesauro é realizada baseando-se no valor discriminatório dos termos, tal como proposta por Salton *et al*. [SYY75]. Esse valor discriminatório indica o quanto um termo é representativo para um documento, de acordo

com a frequência desse termo no documento e de acordo com a frequência do mesmo na coleção de documentos.

Esses valores discriminatórios podem ser de três tipos, onde *n* é a quantidade de documentos na coleção:

- Indiferentes: No caso dos termos em que a frequência de ocorrência nos documentos é menor do que o *n*/100. São considerados termos com baixas frequências.
- Discriminatórios pobres: No caso dos termos cuja frequência é maior do que *n*/10. Esses são considerados termos de alta frequência.
- Discriminatórios bons: Nos casos restantes, ou seja, termos com uma frequência entre *n*/10 e *n*/100.

O tesauro deve ser composto por termos que estão intimamente relacionados na coleção. Para isso, é utilizado um método de agrupamento (*clustering*), onde os termos que contêm um valor discriminatório indiferente, isto é, termos com baixas frequências nos documentos, servirão para criar as classes dos agrupamentos. Os termos restantes dos documentos são agrupados dentro dessas classes. O problema dessa técnica ocorre quando todos os termos têm frequências baixas, ou seja, pouca informação para ser adquirida e, com isso, os agrupamentos podem ficar pouco significativos.

3.1.2. Church e Hanks [CH90]

O trabalho de Church e Hanks [CH90] utiliza a Informação Mútua, que compara a probabilidade de observar os termos t_i e t_j juntos (a probabilidade da junção) com a probabilidade de observar os termos t_i e t_j independentemente.

Sendo $P(t_i)$ e $P(t_j)$ a probabilidade do termo t_i e do termo t_j respectivamente, e $P(t_i,t_j)$ a probabilidade dos termos t_i e t_j aparecerem juntos, define-se a Informação Mútua $IM(t_i, t_j)$ entre os termos t_i e t_j como sendo:

$$IM(t_i, t_j) = log_2 \left(\frac{P(t_i, t_j)}{P(t_i)P(t_j)} \right) \quad (5)$$

Se existe uma associação entre t_i e t_j, então a probabilidade $P(t_i, t_j)$ de aparecerem juntos será muito maior do que a probabilidade de encontrar somente t_i ou somente t_j e, consequentemente, a informação mútua terá um valor $IM(t_i, t_j) >> 0$. Porém, se não existe uma relação significativa entre t_i e t_j, então a probabilidade de encontrar t_i e t_j juntos será

praticamente a mesma de encontrá-los sozinhos $P(t_i, t_j) \approx P(t_i)\ P(t_j)$ e a informação mútua entre eles será $IM(t_i, t_j) \approx 0$. Entretanto, se t_i e t_j raramente aparecem juntos, a probabilidade de ambos será muito menor do que a probabilidade de cada um deles em separado, forçando assim $IM(t_i, t_j) << 0$.

Uma explicação mais simplista é dada por Manning *et al.* [MRS08], e diz que a Informação Mútua mede o quanto um termo nos fala a respeito de outro. Essa medida verifica então a probabilidade de um termo andar sempre com outro.

Embora o trabalho de Church e Hanks não vise à criação automática de tesauros, ele serve de suporte para esta tarefa, pois esse trabalho fornece a base estatística de uma interessante variedade de fenômenos linguísticos. Essa variedade de fenômenos linguísticos permitiria incluir a utilização de termos com “médico” quando uma pessoa é perguntada por um termo semelhante a “enfermeiro”.

3.1.3. Kaji *et al.* [KMAY00]

O trabalho de Kaji *et al*. [KMAY00] utiliza uma abordagem de associação entre as palavras, também conhecida como associação de primeira ordem [Rug92]. Essa abordagem propõe que a similaridade semântica possa ser computada pelo entendimento lexical entre os vizinhos. Por exemplo, a similaridade semântica entre as palavras *vermelho* e *azul* pode ser definida pelo fato de que ambas coocorrem frequentemente com palavras como *cor*, *flor*, *carro*, *escuro*, *claro*, e assim por diante. Assim, para calcular o grau de similaridade entre as palavras que coocorrem, esse trabalho utilizou a medida de Informação Mútua, proposta por Church [CH90].

A criação de tesauro proposta por Kaji *et al*. [KMAY00] consiste na extração de termos, extração de coocorrências dos termos e análise de correlação, como mostrado na Figura 3.1. A seguir no texto é explicado cada passo para a geração automática de tesauros a partir de corpus, segundo estes autores.

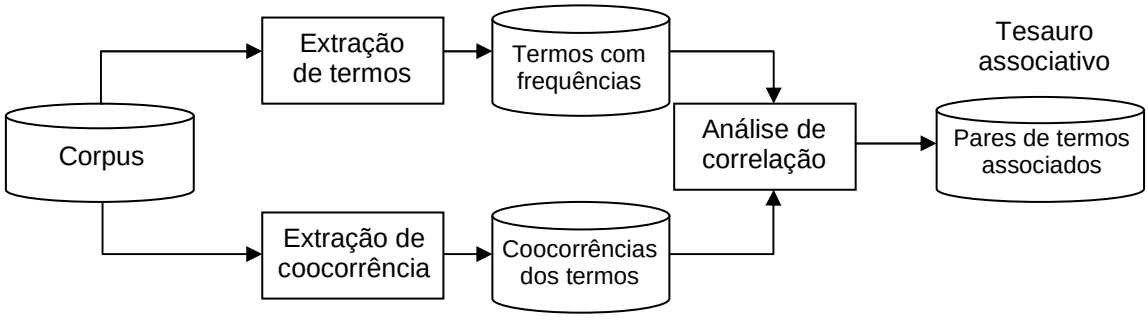

Figura 3.1. Passos para a geração do tesauro (adaptado de [KMAY00])

- Extração de termos

Kaji *et al*. [KMAY00] colocam que os melhores termos para representar conceitos são substantivos, simples ou compostos, que ocorrem frequentemente no corpus. Por isso, são extraídos do texto substantivos simples e compostos em que a frequência de ocorrência ultrapassa um determinado limiar pré-estabelecido. Para a extração também é utilizada uma *stoplist*. Essa *stoplist* é utilizada para fazer uma filtragem no corpus, retirando assim os termos que sabidamente não apresentam interesse nessa relação.

Na extração de termos compostos também é necessária a utilização de uma *stoplist*, já que muitos termos compostos extraídos não constituem um bom termo, pois são combinações de palavras sem significado para o texto. Kaji *et al*. utilizaram uma *stoplist* para o primeiro e uma para o último elemento do termo composto, eliminando assim termos sem significado relevante.

Outro problema encontrado ao se extrair termos compostos é a ambiguidade estrutural dos termos, isto é, termos compostos que podem conter outros termos compostos dentro deles. Para isso, Kaji *et al*. classificaram os termos compostos em não-máximos e máximos. Um termo composto não-máximo é aquele que faz parte de um termo composto maior e um termo composto máximo é aquele que não é parte de um termo composto maior. Assim, para a correta extração de termos compostos não máximos foi desenvolvido um método estatístico de desambiguação estrutural, descrito a seguir.

- Desambiguação estrutural

A regra geral do método de desambiguação estrutural é dada como: se existe um termo composto *CN* que inclui dois termos candidatos máximos CN_1 e CN_2, e esses termos são incompatíveis entre si, então, se um dos dois candidatos é mais frequente, a estrutura de *CN* incluindo a estrutura desse candidato mais frequente é escolhida.

Assim, podemos citar como exemplo o caso de um termo composto por três elementos $W_1W_2W_3$. Esse termo composto pode ser constituído de duas possíveis estruturas $W_1(W_2W_3)$ ou $(W_1W_2)W_3$. Assim é verificado no corpus se o termo composto máximo W_2W_3 ocorre mais vezes do que o termo composto W_1W_2 e então a estrutura $W_1(W_2W_3)$ é preferida. Do contrário, se o termo composto máximo W_1W_2 aparece mais vezes do que o termo W_2W_3, então a estrutura $(W_1W_2)W_3$ é preferida.

Para a contagem de frequências dos termos compostos máximos, Kaji *et al*. utilizaram duas abordagens. São elas: a estatística global das frequências, isto é,

contaram as frequências dos termos compostos do corpus inteiro e essas serviram para desambiguar todos os termos compostos do corpus; e a estatística local das frequências, isto é, contaram as frequências dos termos compostos em cada documento e assim usaram-nas para desambiguar somente os termos do documento correspondente.

A avaliação desses dois métodos mostrou que o método baseado em estatística local obteve melhores resultados do que o método baseado em estatística global para a desambiguação dos termos compostos.

- Extração de coocorrências e análise de correlação

Na extração de coocorrências é coletado qualquer par de termos semanticamente ou contextualmente associados, não importando o tipo de associação. Nesse trabalho foi utilizada uma técnica de janela para a extração.

A técnica de janela extrai pares de termos que ocorrem juntos dentro de uma janela que vai se movendo através do texto. Essa janela é composta por um conjunto de sentenças de *n* palavras, sendo *n* um número previamente escolhido. O tamanho da janela pode ser escolhido arbitrariamente e, devido ao custo computacional, os autores escolheram um tamanho de janela entre 20 e 50 palavras.

Esses pares de palavras ainda são filtrados para que não apareçam pares de substantivos compostos que já foram previamente extraídos, pois se eles fossem incluídos na extração de coocorrência, isso causaria redundância.

No processo de análise de correlação os autores predeterminaram um número máximo de termos associados a cada termo e um limiar para a Informação Mútua, diminuindo assim a quantidade de termos recuperados. Os termos que compõem o tesauro são selecionados com base na ordem decrescente de Informação Mútua.

3.1.4. Nakayama *et al.* [NHN07]

Em [NHN07] e posteriormente em [INHN08, NIHN08] é utilizada a estrutura da Wikipédia para a construção de um tesauro associativo. Esse tesauro é baseado no montante de informações que podem ser obtidas através da análise da estrutura da mesma, principalmente através dos *hyperlinks* contidos em cada um dos seus artigos.

Para analisar a relevância dos *hyperlinks* de cada página e posteriormente encontrar as páginas mais semelhantes, foi desenvolvida uma medida baseada no cálculo de Frequência do Termo – Frequência Inversa do Documento (do inglês, *Term Frequency*

– *Inverse Document Frequency* - TF-IDF), que os autores chamaram de *Path Frequency - Inversed Backward link Frequency* (PF-IBF). Essa medida foi criada com base na ideia que uma página que compartilha *hyperlinks* com uma página específica e não compartilha com outras páginas, tem um alto valor de PF-IBF com esta página específica.

O tesauro é criado verificando-se todas as conexões da Wikipédia e calculando para cada uma delas o valor de PF-IBF. Com isso, os conceitos que têm maior valor de PF-IBF aparecerão associados no tesauro.

3.2 Técnicas baseadas em sintaxe

A técnica aqui descrita necessita, além de métodos estatísticos, também de ter o corpus anotado por um analisador sintático, conhecendo assim as categorias gramaticais dos termos e a função dos mesmos na frase. Assim, podem-se obter relações sintáticas entre termos, como o sujeito, o objeto direto ou o objeto indireto de um verbo, por exemplo.

3.2.1. Grefenstette [Gre94]

Grefenstette [Gre94] descreve a criação de um tesauro a partir de textos, porém baseado em sintaxe, isto é, que utiliza os contextos sintáticos para o cálculo da similaridade entre as palavras para a construção do tesauro. Entende-se por contexto sintático qualquer conjunto de palavras que estabeleçam uma relação sintática com outra palavra no corpus. Para estabelecer esse contexto podem ser identificadas diversas relações sintáticas entre as palavras, como sujeito, objeto etc.

Esse trabalho foi o primeiro a utilizar informações sintáticas e também foi um dos primeiros a encontrar palavras semanticamente similares (baseadas na sintaxe) através de uma busca automática. Ele ainda apresenta o sistema SEXTANT (*Semantic EXtraction from Text via Analysed Networks of Terms*), que utiliza contextos sintáticos para obter a similaridade entre as palavras.

Para a construção do tesauro, Grefenstette faz a extração de determinados contextos sintáticos das palavras e através deles faz a associação entre as palavras utilizando a medida de similaridade de Jaccard ponderada. Como resultado, o autor obtém uma lista de palavras semanticamente relacionadas a cada substantivo do corpus. Nesta técnica são comparados sintagmas nominais, a fim de identificar aqueles que são semanticamente relacionados.

Para a criação do tesauro, Grefenstette realiza os seguintes passos para a obtenção de listas de palavras relacionadas:

a) *Tokenizar* o corpus com uma gramática regular. Os *tokens* são morfologicamente analisados e um léxico provê as categorias (para então chegar-se ao *POS tagging*) de cada *token*;

b) O texto etiquetado passa por um desambiguador estocástico, deixando cada *token* com apenas uma etiqueta, isto é, eliminando as ambiguidades.

c) O texto sem ambiguidades passa por um *parser* que etiqueta os sintagmas nominais e verbais. Esses sintagmas é que permitem identificar se um substantivo está ligado a um verbo como sujeito ou como objeto.

d) São extraídos todos os contextos sintáticos de cada palavra em todas as suas ocorrências no corpus. São extraídos todos os contextos sintáticos em que os substantivos estão relacionados com adjetivos e sintagmas preposicionais que os modificam. Se os substantivos estão conectados a um verbo, são extraídos os contextos sintáticos, identificando se fazem papel de sujeito ou objeto.

e) É utilizada a medida de similaridade para gerar uma lista de palavras semelhantes. A medida é uma variante da medida de Jaccard ([Tan58] apud [Gre94]), que utiliza pesos associados aos contextos sintáticos. As duplas de palavras mais similares, isto é, com uma medida de similaridade mais alta, formarão o tesauro daquele conceito. Assim, a medida de Jaccard ponderada identificará como palavras mais similares aquelas que compartilham um número significativo de contextos semelhantes.

A seguir é explicado o que são os contextos sintáticos e, posteriormente, como é medida a similaridade entre as palavras.

- Contextos sintáticos

Grefenstette extrai os contextos sintáticos das palavras que aparecem em um mesmo sintagma nominal ou entre o núcleo de um sintagma nominal e o núcleo de um sintagma verbal relacionado. Assim, uma relação sintática é denotada por uma tripa:

<R, p1, p2>

Nessa tripla, R é uma relação do tipo:

- ADJ: um adjetivo que modifica um substantivo;
- NN: um substantivo que modifica outro substantivo (especialmente presente na língua inglesa);
- NNPREP: um substantivo que modifica outro substantivo utilizando uma preposição;
- SUBJ: um substantivo que tem o papel de sujeito de um verbo;
- DOBJ: um substantivo que tem o papel de objeto direto de um verbo;
- IOBJ: um substantivo que é objeto indireto de um verbo.

A Tabela 3.1 apresenta alguns exemplos das relações sintáticas extraídas por Grefenstette em seu trabalho.

Tabela 3.1. Exemplos de relações sintáticas (adaptados de [Gre94])

Expressões	Relações sintáticas
Possible causes	<ADJ, *cause*, *possible*>
The cause of neonatal jaundice	<NNPREP, *cause*, *jaundice*>
No cause could be determined	<DOBJ, *determine*, *cause*>
Death cause	<NN, *cause*, *death*>

Assim, para cada palavra no texto são encontradas as relações sintáticas em que ela aparece. Grefenstette não faz diferenciação entre as relações *ADJ*, *NNPREP* e *NN,* porém, quando um substantivo é objeto direto de um verbo, ele é identificado de forma diferente, não misturado aos outros contextos sintáticos para a mesma palavra. Assim, no exemplo da Tabela 3.1, os contextos sintáticos da palavra *cause* seriam: <*possible*>, <*jaundice*>, <*death*> e <*DOBJ, determine*>.

- Medida de similaridade

Para fazer a construção do tesauro é necessária uma medida que permita agrupar as palavras. Grefenstette calcula a similaridade entre duas palavras, pal_m e pal_n em função da quantidade de contextos sintáticos que elas compartilham. Para o cálculo, Grefenstette leva em conta os pesos globais e pesos locais de cada palavra no contexto sintático. O peso global *pg* leva em conta a quantidade de palavras diferentes que estão associadas ao contexto cs_j no corpus e pode ser computado pela fórmula citada também no trabalho de Gasperin [Gas01]:

$$pg(cs_j) = 1 + \frac{\sum_j f(pal_i, cs_j)\left(\log\left(f(pal_i, cs_j)\right) - \log\left(f(cs_j)\right)\right)}{\log(npals)} \over f(cs_j) \quad (6)$$

Onde $f(pal_i, cs_j)$ representa a frequência com que a palavra *i* ocorre no contexto sintático *j*, $f(cs_j)$ representa a frequência desse contexto sintático no corpus, e *npals* representa o número de palavras diferentes no corpus.

O peso local *pl* é baseado na frequência do contexto sintático como modificador de uma palavra e pode ser calculado por:

$$pl(pal_i, cs_j) = \log\left(f(pal_i, cs_j)\right) \quad (7)$$

O peso total *pt* é calculado com a multiplicação dos pesos locais e globais. Assim, a medida de Jaccard ponderado *JP* entre duas palavras pal_m e pal_n é calculada por:

$$JP(pal_m, pal_n) = \frac{\sum_j \min\left(pt(pal_m, cs_j), pt(pal_n, cs_j)\right)}{\sum_j \max\left(pt(pal_m, cs_j), pt(pal_n, cs_j)\right)} \quad (8)$$

Depois de computada a similaridade para todos os contextos do corpus, são geradas listas de palavras mais semelhantes para cada palavra do corpus. Essas listas de palavras mais semelhantes é que compõem o tesauro.

3.2.2. Lin [Lin98]

O artigo descrito em [Lin98] relata a criação de um tesauro baseada na extração de contextos sintáticos para cada termo. É feita, além da extração de contextos sintáticos para substantivos (como era feita por Grefenstette [Gre94]), a extração para adjetivos, verbos e advérbios. Para essa extração, são identificadas triplas no corpus anotado sintaticamente. Essas triplas são compostas de dois termos e a relação gramatical existente entre eles na frase. Como exemplo, da frase "*I have a brown dog*." são extraídas as triplas:

(*have* subj *I*), (*I* subj-of *have*), (*dog* obj-of *have*), (*dog* adj-mod *brown*), (*brown* adj-mod-of *dog*), (*dog* det *a*), (*a* det-of *dog*)

Essas triplas são da forma (*w*,*r*,*w'*), onde *w* e *w'* são termos e *r* é a relação existente entre esses termos. No exemplo citado, temos a relação (*have*, subj, *I*), sendo o termo "*I*" o sujeito do verbo "*have*".

Lin considera não só o sentido direto da relação sintática, mas também o sentido inverso da mesma. Assim, de uma relação de sujeito entre “*I*” e “*have*”, são extraídos os contextos (*have* subj *I*) e (*I* subj-of *have*). Lin acredita que, quanto maior a quantidade de contextos sintáticos de uma palavra, maior a confiabilidade dos resultados da comparação com as demais palavras.

Para a construção do tesauro, Lin propôs uma medida de similaridade que leva em conta o valor de informação contido em uma relação e o valor de informação contido em todas as relações que os termos da primeira compartilham no corpus. As palavras que tiverem os valores da medida de similaridade mais altos com relação a uma dada palavra formarão a lista de palavras mais similares a tal palavra do corpus.

3.2.3. Gasperin [Gas01] e Gasperin e de Lima [GL03]

Gasperin faz a geração automática de um tesauro para a língua portuguesa, criando uma extensão da proposta de Grefenstette [Gre94]. Nessa proposta, a autora integra as seguintes extensões ao trabalho de Grefenstette:

- Leva em conta a bidirecionalidade das relações gramaticais, como foi feito no trabalho de Lin [Lin98];
- Diferencia as relações sintáticas extraídas, não diferenciadas no trabalho de Grefenstette, como *ADJ* e *NNPREP*;
- Adiciona a relação entre substantivos de um mesmo verbo (*SOBJ*), quando um deles era sujeito e o outro objeto desse verbo;
- Leva em conta a preposição quando a mesma está em uma relação entre substantivos. Por exemplo, em uma relação “marca de camisa” ou “marca na camisa”, Grefenstette extrai o mesmo contexto sintático para ambas. Levando em conta a preposição, são extraídos dois contextos sintáticos diferentes.

Para a construção do tesauro, é utilizada a medida de similaridade de Jaccard ponderado como apresentado no trabalho de Grefenstette, obtendo uma lista com os termos mais similares a um determinado termo.

3.3Técnicas baseadas em Análise Semântica Latente

As técnicas descritas a seguir utilizam a Análise Semântica Latente (do inglês, *Latent Semantic Analysis* - LSA), ou adaptações da mesma, para descobrir relações "latentes" entre os termos, isto é, relações que não são aparentes se considerarmos apenas a frequência dos termos nos documentos.

3.3.1. Dumais *et al*. [DFL+88] e Landauer e Dumais [LD97]

O trabalho de Dumais *et al.* [DFL+88], que foi o primeiro a utilizar a aplicação da LSA, inicialmente chamada LSI. Nesse trabalho, Dumais *et al.* aplicaram a técnica de LSA para a recuperação de documentos médicos em um banco de dados composto por 1033 títulos de resumos de documentos médicos.

A recuperação de documentos com a aplicação da LSA foi feita sobre 30 consultas disponíveis no banco de dados. Com a aplicação da LSA, obtiveram uma precisão de 0.51, enquanto que com recuperação por *term matching*, isto é, a recuperação buscando a mesma ocorrência do termo na consulta e no documento, foi de 0.45. Essa melhor precisão na obtenção dos documentos mostrou que a técnica de LSA conseguia encontrar relações entre os termos que eram perdidas pelo *term matching*.

Em 1989 os autores do trabalho [DFL+88] registraram a patente "*Computer information retrieval using latent semantic structure*" referente ao método de recuperação de documentos utilizando a LSA. A partir de 1993 Dumais começou a participar do Text REtrieval Conference (TREC) apresentando os trabalhos [Dum93, Dum94, Dum95] e mostrando a eficiência da técnica de LSA quando aplicada na coleção do TREC.

Em 1997, Landauer e Dumais no trabalho [LD97] buscaram encontrar associações que acontecem na memória humana, como aquela devida à similaridade entre termos. Para identificar esse tipo de associação os autores utilizaram os espaços gerados pela Análise Semântica Latente, visando identificar os termos que ocorriam nas mesmas dimensões espaciais.

A técnica de LSA foi utilizada pelo fato de a mesma associar termos que ocorrem em padrões similares nos documentos, mesmo se eles não coocorrem dentro de um documento. Como teste, Landauer e Dumais utilizaram o corpus do *Test Of English as a Second Language* (TOEFL) que contém 80 questões de múltipla escolha, cada uma delas

contendo 4 alternativas de resposta, sendo uma delas (resposta correta) composta pelos sinônimos.

Após fazer a aplicação da LSA no corpus e encontrar as respostas para as respostas do questionário, Landauer e Dumais observaram que fazendo a indução do conhecimento indiretamente utilizando a coocorrência dos termos em um corpus com uma grande quantidade de documentos (como o corpus do TOEFL), a técnica de LSA adquire um conhecimento sobre o vocabulário de inglês comparável ao conhecimento obtido por estudantes.

3.3.2. Yang e Powers [YP08]

O trabalho de Yang e Powers [YP08] apresenta uma proposta de criação de tesauros que utiliza, além de informações linguísticas, uma adaptação da técnica de LSA. Nesse trabalho Yang e Powers colocam que a maioria dos métodos de construção de tesauros simplesmente ignora a saliência das relações gramaticais e efetivamente fundem-nas em apenas um único contexto.

Yang e Powers colocam que os algoritmos para construção de tesauros normalmente calculam a distribuição das palavras em um vetor de termos e esse cálculo pode ser sintaticamente condicionado (e.g. através de relações gramaticais), como apresentado no trabalho de Grefenstette [Gre94] ou não condicionado (e.g. através da abordagem denominada *bag of words*), como apresentado na abordagem de Kaji *et al*. [KMAY00].

Embora os métodos que utilizam relações gramaticais consigam ter uma qualidade maior nos termos gerados, se comparados a uma abordagem *bag of words*, eles também juntam as dependências sintáticas em um contexto unificado na construção de um tesauro. Dessa maneira, não conseguem mostrar diferenças latentes nas relações gramaticais para a determinação do significado da palavra em um contexto.

Para obter melhores resultados, Yang e Powers propõem primeiramente categorizar os contextos em termos de relações gramaticais, e então utilizar uma adaptação da técnica de LSA para as categorias gramaticais, de forma a obter as relações latentes entre os termos. Como resultado, é obtida uma matriz de termos relacionados semanticamente. Essa matriz pode ser interpretada de forma que o produto do cosseno entre as linhas dessa matriz representa a similaridade entre dois termos.

3.3.2.1Extração de dependências sintáticas

Segundo Yang e Powers, a representação sintática principalmente depende dos seguintes fundamentos:

- O significado de um substantivo depende de seus modificadores, como adjetivos, outros substantivos ou núcleos nominais de sintagmas preposicionais, bem como do papel gramatical de um substantivo como sujeito ou objeto de uma frase;

- O significado de um verbo depende do seu objeto direto, sujeito ou modificador como, por exemplo, o núcleo nominal em um sintagma preposicional.

Com base nesses fundamentos, sabe-se então que as dependências sintáticas podem prover pistas para encontrar o significado de uma palavra dentro de um contexto. Tendo-se uma tupla $<w_i, r, w_j>$, onde w_i e w_j representam duas palavras e r é a relação de dependência, se w_i modifica w_j através de r, então todos os w_j com um modificador r formarão um contexto para w_i, assim como w_i formará para w_j. Assim, essas dependências sintáticas (r) podem se classificar em quatro grupos:

- **RV**: Verbos com seus modificadores, como todos os advérbios e os núcleos nominais de sintagmas preposicionais;
- **AN**: Substantivos com seus modificadores, como adjetivos e pré/pos modificadores;
- **SV**: Sujeitos e seus predicados; e
- **VO**: Predicados e seus objetos.

Para obter essas dependências Yang e Powers utilizaram o *parser* Link Grammar [ST91]. No total existem 107 tipos de conexões detectadas por esse *parser*. Desses 107 tipos, alguns foram extraídos por Yang e Powers e classificados nos seguintes grupos previamente descritos:

RV

E: Verbos e seus advérbios pré-modificadores

EE: Advérbios e seus advérbios pré-modificadores

MV: Verbos e seus pós-modificadores, como advérbios e sintagmas preposicionais

AN

A: Substantivos e seus adjetivos pré-modificadores

AN: Substantivos e seus substantivos pré-modificadores

GN: Substantivos que estão ligados a nomes próprios

M: Substantivos e seus vários pós-modificadores como sintagmas preposicionais, adjetivos e particípios

SV

S: Sujeito-substantivos/gerúndios e seus verbos finitos.

Derivados de S, como Ss*g para gerúndios e os seus predicados, Sp para substantivos plurais etc.

SI: Sujeito e seus verbos quando ocorre a inversão do sujeito e do verbo em questões

VO

O: Verbo e seus objetos diretos e indiretos

OD: Verbos e seus complementos de distância

OT: Verbos e seus objetos de tempo

P: Verbo "ser" conectado às palavras que podem ser seus complementos.

Derivados de *P*, como *Pp* quando o verbo “ser” está conectado com uma preposição, ou *Pa* quando o verbo “ser” está conectado com um adjetivo etc.

Após fazer o processamento do corpus, Yang e Powers extraem os relacionamentos para construir quatro matrizes paralelas ou conjuntos de coocorrência, denominados R_x: RV_x, AN_x, SV_x e Vo_x em termos dos quatro grupos de dependências sintáticas. As linhas dessa matriz são denominadas Rv_x, An_x, Sv_x e Vo_x e as colunas são denominadas rV_x, aN_x, sV_x e vO_x respectivamente.

Consideremos SV_x uma matriz *mxn* que representa as dependências de sujeito e verbo, sendo a linha *m*, ou $\{X_{i,*}\}$, correspondente aos sujeitos condicionados como sujeitos de verbos nas sentenças e a coluna *n*, ou $\{X_{*,j}\}$, os verbos condicionados por substantivos como sujeitos. Na Tabela 3.2 é mostrado um exemplo de como ficará a matriz SV_x, onde as frequências *freq_i_j* representam as frequências de ocorrência do *substantivo_i* com o *verbo_j.*

Tabela 3.2. Exemplo de matriz SVx

	Verbo_1	Verbo_2	Verbo_3
Substantivo_1	freq_1_1	freq_1_2	freq_1_3
Substantivo_2	freq_2_1	freq_2_2	freq_2_3
Substantivo_3	freq_3_1	freq_3_2	freq_3_3

Após a obtenção das matrizes de dependência sintática, o próximo passo para a construção do tesauro é realizar a normalização dos termos. Esse processo é feito substituindo a frequência de cada célula $freq(X_{i,j})$ pelo valor da sua informação, usando $log(freq(X_{i,j})+1)$, diminuindo assim o espaço existente entre eventos muito frequentes e eventos raros. Após, aplica-se uma adaptação da técnica de LSA, para encontrar eventos mais surpreendentes, ao invés de apenas fazer a redução da matriz esparsa obtendo apenas as coocorrências dos termos (tratando como *bag of words*).

Por fim, Yang e Powers procuram pelos termos mais semelhantes contidos na matriz gerada pelo processo de LSA. A similaridade distribucional é utilizada para a comparação dos valores da matriz resultante $\hat{A}$. Para a obtenção dos termos é normalmente utilizado um método de comparação de vetores. A medida do Cosseno obtém o ângulo formado por dois vetores e pode ser definida como:

$$cos\theta = \frac{x \cdot y}{\| x \| \| y \|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \quad (9)$$

Onde ||x|| e ||y|| são a Norma Euclidiana dos vetores *x* e *y* respectivamente, isto é, representam o comprimento dos vetores. Assim, os autores constroem o tesauro utilizando o método adaptado do LSA.

4. RECURSOS, FERRAMENTAS E APLICAÇÃO

Neste capítulo é descrito o processo de criação de uma ferramenta para a geração automática de um tesauro baseado no método de Kaji *et al*. [KMAY00], dois tesauros baseados no método desenvolvido por Grefenstette [Gre94] e dois tesauros baseados no método apresentado por Yang e Powers [YP08]. O capítulo começa apresentando o corpus utilizado e, a seguir, o modo como foi realizada a criação da ferramenta que implementa os três métodos descritos.

4.1 O corpus

Para a geração dos tesauros associativos foram escolhidas abordagens que utilizam um corpus para a identificação dos termos relacionados. Para a criação desse corpus, foram coletados documentos do domínio legal e de acesso público.

Esses documentos foram obtidos de *sites* governamentais que disponibilizam leis, normas ou *guidelines* do domínio legal. Todos os documentos foram obtidos com versões em língua inglesa das normas, sendo aproximadamente metade dos documentos obtidos de normas dos Estados Unidos. Austrália, Nova Zelândia, Reino Unido e Canadá são outros países com grande porcentagem de documentos, restando aproximadamente 20% dos documentos referentes a países que não têm a língua inglesa como idioma oficial. Embora países como Espanha, Coréia, China e Japão não tenham a língua inglesa como idioma oficial, os documentos recuperados são traduções oficiais das leis desses países.

A partir dessas fontes, foi coletado um total de cem documentos, sendo o maior documento, uma lei americana contendo 72 mil palavras, e o menor documento, uma lei australiana, contendo 128 palavras. A coleção de documentos é composta por um total de 1.122.836 palavras. A Tabela 4.1 apresenta algumas estatísticas sobre o corpus utilizado.

Tabela 4.1. Estatísticas referentes ao corpus

Elemento	**Ocorrências**	**Quantidade (sem repetição)**
Substantivo	665.790	8.367
Verbo	235.266	4.808
Adjetivo	137.916	3.199
Preposição	251.967	119
TOTAL	1.290.939	16.493

Uma dificuldade encontrada ao lidar com textos do domínio legal refere-se à estrutura do texto no documento, composta de itens e subitens, tabelas, parágrafos etc. Para um melhor aproveitamento pelo analisador sintático, os textos que compõem o corpus foram previamente tratados. Neste tratamento foram retiradas tabelas, referências a figuras, marcações de itens e caracteres especiais como, por exemplo, o símbolo de parágrafo (§). O Fragmento 3 apresenta um recorte da lei americana "*Unfair or deceptive acts or practices*" no formato original e o Fragmento 4 apresenta a mesma lei após o tratamento.

Fragmento 3. Recorte da lei sem tratamento

Unfair Acts or Practices

- Assessing whether an act or practice is unfair
- An act or practice is unfair where it (1) causes or is likely to cause substantial injury to consumers, (2) cannot be reasonably avoided by consumers, and (3) is not outweighed by countervailing benefits to consumers or to competition. Public policy may also be considered in the analysis of whether a particular act or practice is unfair. Each of these elements is discussed further below.

- The act or practice must cause or be likely to cause substantial injury to consumers.

To be unfair, an act or practice must cause or be likely to cause substantial injury to consumers.

Fragmento 4. Recorte da lei após o tratamento

Unfair Acts or Practices.
Assessing whether an act or practice is unfair.
An act or practice is unfair where it causes or is likely to cause substantial injury to consumers, cannot be reasonably avoided by consumers, and is not outweighed by countervailing benefits to consumers or to competition. Public policy may also be considered in the analysis of whether a particular act or practice is unfair. Each of these elements is discussed further below.
The act or practice must cause or be likely to cause substantial injury to consumers.
To be unfair, an act or practice must cause or be likely to cause substantial injury to consumers.

Para a criação dos tesauros baseados em sintaxe e dos tesauros baseados em Análise Semântica Latente, os textos do corpus foram analisados pelo analisador sintático desenvolvido em Stanford[4] [KM03]. Este analisador sintático utiliza Gramáticas livres de contexto probabilísticas (do Inglês, *Probabilistic Context-Free Grammars – PCFG*) para fornecer uma representação gramatical das relações entre palavras em uma sentença.

Este analisador sintático foi escolhido por ter um alto desempenho com textos em língua inglesa, se comparado com outros analisadores para a mesma língua, como os apresentados no trabalho de Mollá e Hutchinson [MH03]. Segundo Klein e Manning em

[4] *http://nlp.stanford.edu/software/lex-parser.shtml*

[KM03], este analisador sintático tem uma taxa de acerto de 86,36%. Infelizmente até a versão utilizada do analisador sintático (versão 1.6.2), o mesmo não oferecia a opção de lematização dos termos do corpus.

O Fragmento 5 apresenta um trecho do corpus depois de analisado sintaticamente, onde estão em negrito as etiquetas utilizadas para identificar os termos e as relações entre os mesmos, e em itálico as dependências entre os termos.

Fragmento 5. Trecho do corpus após analisado sintaticamente

```
(ROOT
  (S
    (NP (NNS Customers))
    (VP (MD must)
      (VP (VB provide)
        (NP (NN consent))
        (SBAR (IN before)
          (S
            (NP (DT any) (JJ personal) (NN information))
            (VP (VBZ is)
              (VP (VBN transferred)
                (PP (IN from)
                  (NP (PRP$ their) (NN computer)))))))))
    (. .)))

nsubj(provide-3, Customers-1)
aux(provide-3, must-2)
dobj(provide-3, consent-4)
mark(transferred-10, before-5)
det(information-8, any-6)
amod(information-8, personal-7)
nsubjpass(transferred-10, information-8)
auxpass(transferred-10, is-9)
advcl(provide-3, transferred-10)
prep(transferred-10, from-11)
poss(computer-13, their-12)
pobj(from-11, computer-13)
```

Inicialmente o analisador sintático cria a estrutura em árvore da sentença, etiquetando, além da classe gramatical do termo, também as relações sintáticas do termo na frase. Dessa forma obtemos, do exemplo dado no Fragmento 5, o termo "*Customer*", que está marcado como "*NNS*". Este termo é um substantivo, marcado com a etiqueta "*NN*" e seguido de "*S*" significando que o mesmo está no plural. Ainda dessa forma, outras categorias morfossintáticas podem ser observadas como, por exemplo, a etiqueta "*JJ*" para adjetivos, a etiqueta "*VB*" para verbos etc. As etiquetas dos termos seguem o padrão

Penn Treebank e podem ser encontradas no Anexo A deste trabalho. Mais detalhes a respeito das etiquetas podem ser encontradas no trabalho de Marcus *et al*. [MKM+94].

A segunda fase da análise de cada frase é composta pela identificação das relações entre os termos na frase. Nessa fase são marcadas relações entre adjetivos e substantivos, substantivos e outros substantivos, verbos com seu sujeito e objetos direto e indireto etc. Assim, observando o exemplo dado no Fragmento 5, a relação "*nsubj(provide-3, Customers-1)*" pode ser entendida como: o termo "*Customers*" (que ocupa a primeira posição na frase) é o predicativo do sujeito para o verbo "*provide*" (que ocupa a terceira posição na frase). Dessa forma, outras relações entre termos podem ser observadas, como "*dobj()*" referenciando o objeto direto de um verbo, "*amod()*" fazendo a relação entre um adjetivo e um substantivo etc. A forma como são extraídas as dependências e o significado de cada dependência podem ser encontrados no trabalho de Marneffe *et al*. [MMM06].

Para a extração dos contextos sintáticos no processo de geração automática dos tesauros, foram identificados no texto marcado pelo analisador sintático as etiquetas morfossintáticas e as etiquetas das funções sintáticas das palavras. Estas etiquetas são apresentadas na Tabela 4.2.

Tabela 4.2. Etiquetas morfossintáticas identificadas para extração

Categoria morfossintática	**Descrição**
NN[*]	Substantivos
JJ	Adjetivos
IN	Preposições
Relações sintáticas	**Descrição**
nn	Substantivo que modifica um termo
amod	Adjetivo que modifica um termo
_of	Termos modificados através da preposição "of"
[*]subj[*]	Termo que é sujeito de um verbo
iobj	Termo que é objeto indireto de um verbo
dobj	Termo que é objeto direto de um verbo
agent	Termo que é executa a ação de um verbo

Na Tabela 4.2 a cadeia "[*]" significa "quaisquer caracteres". Como exemplo, são extraídos, além de termos marcados como "*NN*" (substantivos no singular), também "*NNS*" (substantivos no plural), "*NNP*" (nome próprio no singular) e "*NNPS*" (nome próprio no plural). Para relações sintáticas, são extraídos "*nsubj*" (sujeito nominal), "*subj*" (sujeito) e "*nsubjpass*" (sujeito na voz passiva).

Como pode ser observado, o analisador sintático utilizado fornece as etiquetas para os termos e para as relações entre termos, porém não fornece a forma canônica dos termos, isto é, a forma como o termo é, sem flexões de gênero, número ou grau. Como o analisador sintático não faz esta diferenciação entre os termos, ao fazer a extração dos contextos sintáticos serão tratados como dois termos diferentes os termos que aparecerem nas formas plural e singular, por exemplo.

Na geração do tesauro esse tipo de diferenciação trouxe resultados positivos e negativos. Por um lado, foram observados termos tratados diferentemente apenas pela declinação de número, o que poderia ter aumentado a quantidade de relações do termo-chave com outros termos relacionados, caso houvesse essa normalização. Por outro lado, foi observado que, na maioria dos casos, quando existia o plural do termo-chave no corpus, ele aparecia como o termo mais similar ao termo-chave. Essa aparição tende a confirmar a correção do método de geração automática do tesauro, visto que o termo-chave no plural tende a compartilhar os mesmos contextos sintáticos que o termo-chave no singular.

Nas seções seguintes serão descritas as etapas para a criação do tesauro em cada um dos três métodos descritos anteriormente. O primeiro é o tesauro gerado pelo método de Kaji *et al.* [KMAY00], o segundo tesauro é gerado pelo método de Grefenstette [Gre94] e por fim, o terceiro tesauro é gerado pelo método de Yang e Powers [YP08].

4.2 Tesauro baseado no trabalho de Kaji *et al*. [KMAY00]

Para a criação do tesauro baseado no trabalho de Kaji *et al*. [KMAY00], denominado T1, foram criadas funcionalidades para atender às etapas específicas de: extração dos termos compostos do corpus, através do processo de desambiguação estrutural, conforme exposto na subseção 3.1.3, e para o gerenciamento das ferramentas utilizadas para o processo de extração de termos e cálculo de Informação Mútua entre os termos.

Para a criação das funcionalidades foi utilizada a linguagem de programação PERL [WCS96]. A estrutura completa das funcionalidades criadas é apresentada na Figura 4.1, onde as setas tracejadas indicam leitura ou gravação em arquivos e as setas simples indicam o fluxo do processo. Cada uma das etapas é explicada detalhadamente a seguir.

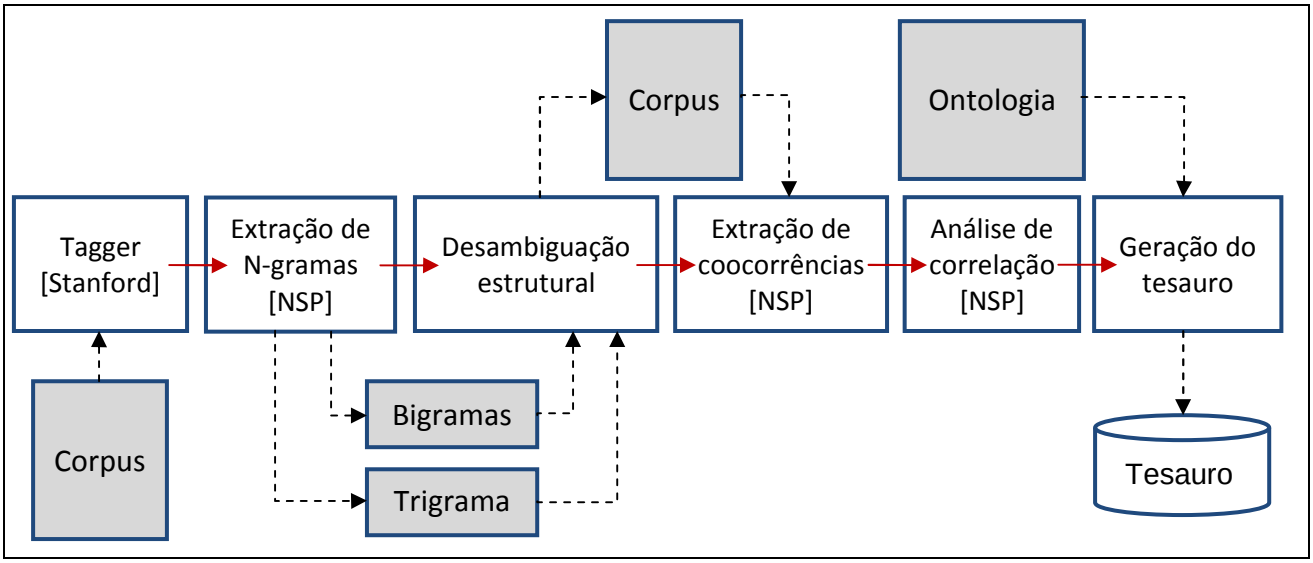

Figura 4.1. Estrutura para a criação do tesauro T1

4.2.1. Etapa 1: Extração de n-gramas e desambiguação estrutural

Nesta etapa são extraídos os termos compostos por duas palavras (bigramas) e por três palavras (trigramas). Esse tipo de extração é feito por uma ferramenta desenvolvida por Banerjee e Pedersen chamada Ngram Statistical Package (NSP)[5]. Detalhes sobre a ferramenta NSP podem ser encontrados no trabalho de Banerjee e Pedersen [BP03].

Depois de extraídos todos os bigramas e trigramas foi criada uma funcionalidade para verificar qual estrutura era mais frequente usando a mesma proposta descrita no trabalho de Kaji *et al.* [KMAY00]. O resultado dessa extração é uma lista contendo bigramas e trigramas que são marcados no corpus original como termos que contêm uma palavra apenas (unigrama). Isso é feito através da união das palavras com o caractere sublinha "_". Por exemplo, o termo "*personal information*" foi identificado no corpus como sendo um bigrama, logo ele é substituído pelo termo "*personal_information*" que passa a ser identificado como um unigrama.

4.2.2. Etapa 2: Extração de coocorrências

O processo de extração de coocorrências utiliza a ferramenta NSP para fazer a identificação de termos que coocorrem dentro de uma janela. A ferramenta permite ao usuário selecionar o tamanho da janela para a extração dos termos que coocorrem. A extração das coocorrências foi executada para uma janela contendo 30 termos, pois compreende o tamanho adotado por Kaji *et al.* [KMAY00].

O resultado da extração de coocorrências é um arquivo texto contendo todas as coocorrências contidas no corpus dentro de uma janela de 30 termos, com a frequência com que aparecem essas coocorrências, seguida da frequência com que o primeiro termo aparece junto a outros termos na primeira posição do bigrama, seguido da frequência com que o segundo termo aparece junto a outros termos na segunda posição do bigrama. O Fragmento 6 apresenta um trecho da lista de coocorrências extraídas do corpus com suas respectivas frequências.

Fragmento 6. Exemplos de extração de coocorrências do corpus

```
...
personal_data<>data<>747 28148 43793
processing<>personal_data<>716 21573 28196
subsection<>person<>662 45969 60090
processing<>data<>604 21573 43793
...
```

4.2.3. Etapa 3: Análise de correlação e geração do tesauro

Esta etapa utiliza a ferramenta NSP para fazer o cálculo da Informação Mútua (IM) entre os termos que coocorrem no corpus. Este cálculo é apresentado na subseção 3.1.2.

Como resultado do cálculo, a ferramenta gera uma lista com os termos relacionados, a posição do termo de acordo com o valor de IM entre os demais termos, a frequência de ocorrência do bigrama, a frequência do primeiro termo acompanhado de outros termos, estando na primeira posição do bigrama, e a frequência do segundo termo quando acompanhado de outros termos, estando na segunda posição do bigrama. O Fragmento 7 apresenta um trecho do resultado da aplicação da IM para o termo "*personal_information*", apresentando os termos relacionados com maior valor de IM.

Fragmento 7. Exemplo de valores de IM para o termo "*personal_information*"

N-gram	**Rank**	**Mutual Information**	**Frequency**
personal_information<>ibm_web_site	47657	0.0000026368	2 53 144
personal_information<>ibm	52820	0.0000021198	2 53 520
personal_information<>personal	57500	0.0000016518	4 53 27804
personal_information<>variety_of_situations	58028	0.0000015990	1 53 18
…			

[5] *http://www.d.umn.edu/~tpederse/nsp.html*

A partir da lista gerada com o valor da IM de cada par de termos, foi criado um programa para fazer a extração dos termos e o valor da IM para cada termo-chave dado como entrada, gerando o tesauro T1. Foi gerado um arquivo no formato XML com dez termos relacionados para cada termo-chave de entrada. O arquivo XML foi criado para fazer a aplicação do tesauro em uma ferramenta de recuperação de informações. Os termos relacionados do tesauro são ordenados no arquivo XML com base na forma decrescente do valor de IM. O Fragmento 8 apresenta um trecho do arquivo XML gerado para a ferramenta de visualização.

Fragmento 8. Trecho do arquivo XML para a ferramenta de visualização

```
…
<seed id="9" term_id="23" term_name="personal_information" type="concept">
    <term id="1" display="ON" similarity="0.0000026368">ibm_web_site</term>
    <term id="2" display="ON" similarity="0.0000021198">ibm</term>
    <term id="3" display="ON" similarity="0.0000016518">personal</term>
    <term id="4" display="ON" similarity="0.0000015990">variety_of_situation</term>
    <term id="5" display="ON" similarity="0.0000015878">rediscIose</term>
    <term id="6" display="ON" similarity="0.0000015878">period</term>
    <term id="7" display="ON" similarity="0.0000014514">resell</term>
    <term id="8" display="ON" similarity="0.0000013637">authorized_recipient</term>
    <term id="9" display="ON" similarity="0.0000012381">paragraph</term>
    <term id="10" display="ON" similarity="0.0000011067">correspond</term>
</seed>
…
```

4.3 Tesauro baseado no trabalho de Grefenstette [Gre94]

Assim como no processo de criação do tesauro pelo método de Kaji *et al*. [KMAY00], a criação dos tesauros pelo método de Grefenstette [Gre94] também exigiu o desenvolvimento de programas na linguagem de programação PERL. Os programas desenvolvidos fazem a extração das relações entre termos no texto, e a geração do arquivo em XML para a visualização dos tesauros.

Foram gerados dois tesauros utilizando o método de Grefenstette [Gre94], pois um deles, denominado T2, utiliza um corte nos contextos sintáticos extraídos (o mesmo corte é utilizado para a geração dos tesauros baseados na adaptação da técnica de LSA). O outro tesauro gerado, denominado T3, não utiliza esse corte de contextos, utilizando, dessa forma, todos os contextos sintáticos extraídos do corpus.

Com os tesauros T2 e T3 gerados, podemos verificar como o corte de contextos pode afetar a criação do tesauro baseado no trabalho de Grefenstette [Gre94] e como este corte pode influenciar os resultados do tesauro gerado utilizando a adaptação da técnica de LSA.

Também podemos fazer uma comparação entre os termos que aparecem no tesauro T3, não aparecem no T2 e que aparecem no tesauro com a adaptação da técnica de LSA, mostrando assim que a adaptação da técnica de LSA pode encontrar termos que antes do corte de contextos estavam semanticamente relacionados. As etapas para a geração destes tipos de tesauros são apresentadas na Figura 4.2 e explicadas a seguir.

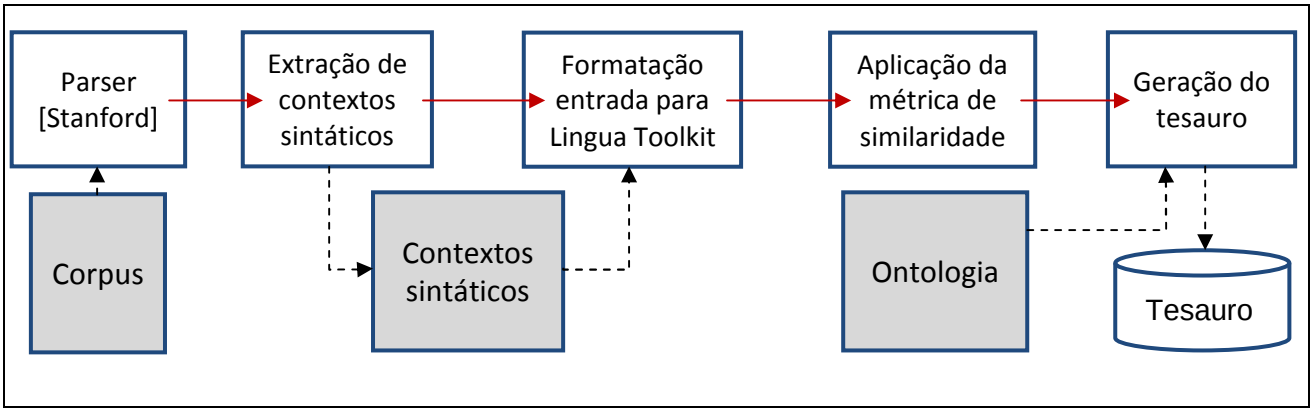

Figura 4.2. Estrutura para criação dos tesauros T2 e T3

4.3.1 Etapa 1: Extração das relações entre termos

Após o corpus ser analisado sintaticamente, são extraídas as relações identificadas pelo analisador sintático. Para tanto, foi criado um programa que faz a identificação de relações para substantivos e de relações que envolvem verbos. A diferença entre elas é que as relações para substantivos ocorrem na modificação de um substantivo, seja por outro substantivo, por um adjetivo, ou por outro substantivo através de uma preposição já nas relações que envolvem verbos, o verbo modifica um sintagma nominal e não apenas um substantivo. Por isso, são extraídos sintagmas nominais em que o núcleo do substantivo é o substantivo identificado na relação sintática pelo analisador sintático. Assim, são extraídos sintagmas nominais onde o núcleo é um substantivo que tem uma relação de sujeito, objeto direto, ou objeto indireto com algum verbo.

Como resultado dessa extração é gerada uma lista contendo todas as relações entre termos identificadas pelo analisador sintático e suas respectivas frequências. O Fragmento 9 apresenta um exemplo da extração dos contextos sintáticos, identificados conforme o trabalho de Grefenstette [Gre94] e explicados na subseção 3.2.1.

Fragmento 9. Exemplos de contextos sintáticos extraídos do corpus

```
…
<ADJ, records, electronic>23
<ADJ, data, personal>1116
<ADJ, information, personal>849
<NN, data, location>32
<NN, privacy, health>29
<NN, european, parliament>41
<NNPREP, section, paragraph>18
<SUBJ, recorded, information >6
<SUBJ, obtained, personal_information >5
<SUBJ, rules, health_information_privacy>21
<OBJ, report, submit>20
<IOBJ,give, individual_information>2
…
```

4.3.2 Etapa 2: Corte de contextos

Para a geração do tesauro T2 é realizado um corte nos contextos sintáticos. Este corte é realizado devido ao mesmo também ser utilizado para a geração dos tesauros baseados no trabalho de Yang e Powers [YP08].

No processo de corte de contextos foram removidos contextos sintáticos que apareciam apenas uma vez no corpus. Dessa forma, por exemplo, caso o contexto sintático "*<ADJ, records, electronic>1*" existisse, ele seria excluído da lista de contextos por ocorrer apenas uma vez no corpus.

4.3.3Etapa 3: Formatação dos dados e geração do tesauro

Para o cálculo de similaridade entre os termos é utilizada uma ferramenta desenvolvida por Pablo Gamallo Otero chamada Lingua Toolkit[6]. Esta ferramenta integra um analisador sintático e um gerador de tesauros. Esse gerador de tesauro utiliza um método adaptado de Grefenstette [Gre94] que faz a geração do tesauro baseado em termos compostos por apenas uma palavra.

Como o gerador de tesauro não faz a obtenção dos termos relacionados baseado em sintagmas, o mesmo não foi utilizado nesta dissertação. Mesmo não utilizando a geração de tesauros proposta pela ferramenta, ela se torna bastante útil ao prover o cálculo de similaridade entre os termos no processo de geração do tesauro. Assim, parte das funcionalidades da ferramenta é utilizada neste trabalho, realizando o cálculo da similaridade entre os termos-chave e os termos relacionados.

[6] *http://gramatica.usc.es/~gamallo/thesaurus/index.htm*

Para o uso das medidas de similaridade da ferramenta Lingua Toolkit, os dados devem ser formatados de modo que a ferramenta processe-os corretamente. Para isso é necessário criar um arquivo com as relações entre todos os termos-chave e os termos relacionados entre os quais se deseja descobrir o valor de similaridade. Além deste arquivo, o arquivo contendo os contextos sintáticos com suas respectivas frequências deve ser passado como parâmetro para a ferramenta.

Foram formatados dois arquivos para serem usados como entrada para a ferramenta Lingua Toolkit, um para a geração do tesauro T2 e o outro para a geração do tesauro T3. Estes arquivos continham os contextos sintáticos e suas respectivas frequências.

Ao fim do processo de obtenção da similaridade entre os termos, foram gerados dois arquivos contendo os termos-chave, os termos relacionados e onze valores de similaridade calculados com base em onze medidas de similaridade diferentes. Um desses arquivos continha os dados para a geração do tesauro T2 e o outro continha os dados para a geração do tesauro T3.

Para a geração de cada um dos tesauros foram extraídos dessas listas os termos-chave com seus respectivos termos relacionados e o valor de similaridade de Jaccard. Os termos foram ordenados decrescentemente pelo valor de similaridade. O Fragmento 10 apresenta um trecho do tesauro T3, contendo o termo-chave "*personal_information*", os termos relacionados e os valores de similaridade para cada um dos termos relacionados.

Fragmento 10. Similaridade gerada para o termo "*personal_information*"

Termo-chave: *personal_information*		
Relacionados:	*health_information*	0.112583
	information	0.111111
	protected_health_information	0.084951
	non_public_personal_information	0.075171
	pii	0.068345
	data	0.053139
	credit_information	0.047382
	patient_safety_work_product	0.044226
	goods	0.041758
	finantial_institution	0.041754
	...	

4.4 Tesauro baseado no trabalho de Yang e Powers [YP08]

A geração do tesauro baseado no método de Yang e Powers [YP08] faz uso de uma adaptação da técnica de Análise Semântica Latente (LSA) para a descoberta de

relações não aparentes entre os termos. Esta técnica utiliza a Decomposição em Valores Singulares (SVD) para realizar os cálculos.

Este método é muito semelhante ao método desenvolvido por Grefenstette [Gre94], porém ao invés de calcular os valores de similaridade utilizando o valor da frequência dos contextos sintáticos, usa um valor semântico obtido pela SVD. Foram gerados dois tesauros utilizando este método. O primeiro deles, denominado T4, utiliza a métrica de similaridade do Cosseno, conforme descrito no trabalho de Yang e Powers [YP08]. O outro tesauro, denominado T5, utiliza a métrica de similaridade de Jaccard, permitindo assim a comparação deste com o tesauro T2, visto que no tesauro T5 somente é adicionada a adaptação da técnica de LSA antes de computar a similaridade entre os termos. As etapas para a geração dos tesauros são apresentadas na Figura 4.3 e descritas a seguir.

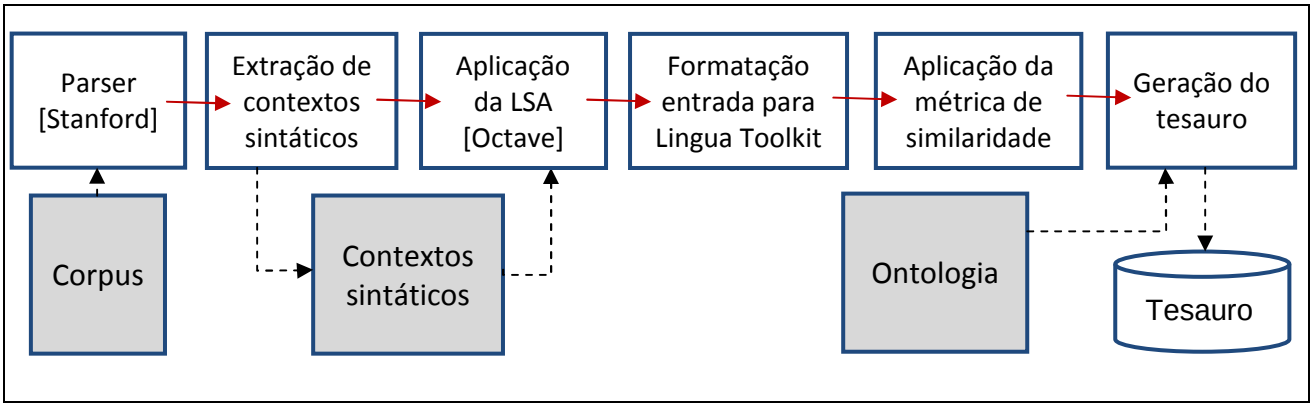

Figura 4.3. Estrutura para criação dos tesauros T4 e T5

4.4.1. Etapa 1: Extração dos contextos sintáticos

A extração dos contextos sintáticos é realizada praticamente da mesma forma neste trabalho e no trabalho de Grefenstette [Gre94]. A diferença está na separação entre os contextos extraídos para a geração das matrizes. No trabalho de Yang e Powers [YP08], são criadas três matrizes *AN*, *SV* e *VO*, tal como exposto na subseção 3.3.2.1.

A primeira matriz contém as relações entre substantivos e substantivos, adjetivos e substantivos, e substantivos e outros substantivos que são modificados através de uma preposição. A segunda matriz contém as relações de verbos com substantivos, quando estes últimos atuam como sujeitos do verbo. A terceira e última matriz contém as relações entre verbos e substantivos, quando estes últimos atuam como objeto (direto ou indireto) dos verbos.

4.4.2. Etapa 2: Geração das matrizes *AN*, *SV* e *VO*

Com os contextos extraídos, foi criada uma matriz *AN*, na forma *An* x *aN*, onde *An* representa os modificadores dos substantivos e *aN* representa os substantivos. (para maiores detalhes ver subseção 3.3.2.1. Foi criada uma matriz *SV* na forma *Sv* x *sV*, onde *Sv* representa os sintagmas nominais quando são sujeitos e *sV* representa os verbos que se relacionam com os sujeitos. Por fim, foi criada a matriz *VO*, na forma *Vo* x *vO*, onde *Vo* representa os verbos que modificam os sintagmas nominais quando os mesmos são objetos da frase, e *vO* representa os sintagmas nominais quando são os objetos desses verbos.

Para uma redução na dimensão das matrizes geradas, foram removidos contextos sintáticos que apareciam apenas uma vez, reduzindo dessa forma o custo computacional para o processamento das matrizes. A Tabela 4.3 apresenta as dimensões de cada uma das matrizes geradas antes da redução e após a redução dos contextos sintáticos. Como pode ser observado, houve uma grande redução no tamanho das matrizes *SV* e *VO*, indicando que muitos sujeitos apareciam apenas uma vez relacionados a um certo verbo no corpus.

Tabela 4.3. Dimensões das matrizes AN, SV e VO

	Matriz sem redução		**Matriz com redução**	
Matriz	**Linhas**	**Colunas**	**Linhas**	**Colunas**
AN	4.349	6.024	2.271	3.000
SV	8.937	3.682	2.230	1.506
VO	7.727	2.667	1.992	1.056

4.4.3. Etapa 3: Decomposição em Valores Singulares

Foi criado um *script* para a interação das matrizes com o *software* Octave[7]. Octave, também conhecido como GNU-Octave, é um *software* livre desenvolvido para a computação matemática. Maiores informações a respeito do Octave e suas funções podem ser encontradas no livro de Quarteroni *et al.* [QGS06].

Octave é programa responsável por fazer a Decomposição em Valores Singulares (SVD) das matrizes *AN*, *SV* e *VO*. Estas matrizes foram decompostas nas matrizes U_{txn}, Σ_{nxn} e V^T_{nxd}, conforme apresentado na subseção 2.2.2. Após a decomposição das matrizes, foi escolhida uma redução para 250 espaços dimensionais, isto é, foram mantidos os primeiros 250 valores singulares da matriz Σ. Segundo Yang e Powers

[7] *http://www.gnu.org/software/octave/*

[YP08] os 20 primeiros valores singulares da matriz Σ representam aproximadamente 50% da variação dos valores da matriz e os primeiros 250 valores singulares representam aproximadamente 75% dessa variação.

As matrizes *AN*, *SV* e *VO* são remontadas utilizando o Octave, que realiza a multiplicação das matrizes U_{txn}, Σ_{nxn} e V^T_{nxd}, porém empregando apenas os 250 primeiros valores singulares na matriz Σ. Como resultado obtêm-se as matrizes AN, SV e VO, porém com os valores semânticos de similaridade entre os termos. Esses valores variam de acordo com os agrupamentos, tendendo a ficarem valores próximos entre termos similares.

4.4.4. Etapa 4: Formatação dos dados para a geração do tesauro

Esta etapa consiste na desconstrução da matriz gerada pelo Octave, de forma a recriar os contextos sintáticos, porém com os valores gerados pela decomposição da matriz ao invés da frequência de ocorrência dos contextos. Após a reconstrução, o processo segue a etapa 2 do processo de construção de tesauros, conforme descrito no trabalho de Grefenstette [Gre94], apresentado na subseção 4.3.2 gerando ao final um tesauro com os termos ordenados por similaridade com o termo-chave.

A ferramenta Lingua Toolkit, que aplica a métrica de similaridade nos vetores de termos gerados, apresenta onze opções de medidas de similaridade. Foram escolhidas duas métricas de similaridade para comparação. A primeira delas (Medida do Cosseno) faz a geração do tesauro T4 e foi escolhida por estar descrita no trabalho de Yang e Powers [YP08]. A outra medida (Jaccard) faz a geração do tesauro T5 e foi escolhida por ser utilizada no trabalho de Grefenstette [Gre94].

5. PROCESSO DE AVALIAÇÃO

Os tesauros gerados com os três diferentes processos adotados nessa dissertação sofreram avaliação, com o objetivo de verificar qual das técnicas apresentadas obtém o melhor resultado. Para a avaliação, cada tesauro foi gerado contendo dez termos-chave e para cada termo-chave foram selecionados os dez termos relacionados mais similares, segundo suas medidas de similaridade.

Por se tratar de um domínio específico, técnicas de avaliação utilizando tesauros construídos manualmente, como a comparação com o tesauro de Roget, o tesauro de Macquarie, o Webster's 7th Dictionary [Gre94], ou a WordNet [YP05, YP08] não puderam ser aplicadas. Optou-se por fazer a avaliação com especialistas do domínio de privacidade de dados. O perfil de cada um dos avaliadores é apresentado no Apêndice A. O processo de escolha de termos-chave, detalhes do sistema de Recuperação de Informações utilizado e os resultados obtidos são apresentados a seguir.

5.1 Escolha dos termos para avaliação

Os termos-chave para a avaliação dos tesauros foram selecionados de uma ontologia do domínio legal em privacidade de dados, conforme apresentado no trabalho de Vieira *et al*. [VSS+10], e de um glossário, também do domínio de privacidade de dados, próprio da empresa parceira do projeto. Reunindo os termos do glossário e os conceitos da ontologia, somávamos um total de 395 termos.

Uma avaliação manual dessa quantidade de termos exigiria muito tempo e tornaria inviável o trabalho devido ao cronograma do mestrado. Decidiu-se então diminuir a quantidade de termos-chave, bem como limitar a quantidade de termos relacionados para cada termo-chave. Em conjunto com os avaliadores, decidiu-se pela escolha de 10 termos-chave para a avaliação.

Para a escolha dos termos a serem selecionados, decidiu-se observar a frequência dos mesmos, visto que, quanto maior a quantidade de ocorrências do termo no corpus, mais relações com outros termos ele tem e, mais significativos serão os termos similares. Em contrapartida, termos mais específicos costumam ter frequência menor.

Optou-se por calcular um limiar que suprisse ambos os casos, não perdendo muitos termos específicos e não deixando uma grande quantidade de termos para serem avaliados. Decidiu-se utilizar um limiar de 50 ocorrências do termo no corpus, pois foi

observado que, com o aumento desse valor de limiar, o número de termos diminuía, porém muitos termos específicos passavam a não aparecer mais na lista. Dessa forma, os termos que tinham uma frequência menor que o limiar não entraram para a lista de termos candidatos a termos-chave. Este valor de limiar reduziu a lista que continha 395 termos inicialmente, para 99 termos.

A partir desta lista de 99 termos, optou-se por fazer uma limpeza manual, retirando termos que não eram do domínio específico. Dessa forma, termos como “*data*”, “*service*”, “*access*”, “*processing*”, “*system*”, “*action*” etc. foram removidos.

Com a retirada de sem significado específico para o domínio de privacidade, restou uma lista com 35 termos. Destes, os avaliadores escolheram 10 termos para servirem de termos-chave para a geração do tesauro. A lista de termos-chave escolhidos é apresentada na Figura 5.1.

• children	• consent	• customer	• data_protection	• data_subject
• marketing	• notice	• personal_data	• personal_information	• regulation

Figura 5.1. Termos-chave escolhidos para serem avaliados

5.2 Geração dos termos relacionados

Para cada um dos termos da lista dos 10 termos-chave, foram gerados 50 termos relacionados, referentes aos métodos empregados. Desses 50 termos, 10 foram obtidos do tesauro construído com método de Kaji *et al*. [KMAY00], gerando o tesauro denominado T1, 10 foram obtidos do tesauro construído com o método de Grefenstette [Gre94] com o corte de contextos, gerando o tesauro T2, 10 foram obtidos do tesauro construído com o método de Grefenstette [Gre94] sem o corte de contextos, gerando o tesauro T3, 10 foram obtidos do tesauro construído com o método de Yang e Powers [YP08] e a fórmula de similaridade do Cosseno, gerando o tesauro T4, e 10 foram obtidos do tesauro construído com o método de Yang e Powers [YP08] e a fórmula de similaridade de Jaccard, gerando o tesauro T5.

Observa-se que foram gerados dois tesauros utilizando o método de Grefenstette, pois o primeiro tesauro gerado (T2, com redução de contextos) serviu de comparação para o método utilizado por Yang e Powers [YP08], que utiliza a mesma redução de contextos utilizada para computar a matriz SVD. Dessa forma, procura-se observar o

ganho que a adaptação da técnica de LSA teria se comparada com uma técnica sem a aplicação da mesma.

O segundo tesauro gerado por essa técnica (T3, tesauro sem redução de contextos) representa a forma tradicional de geração utilizada por Grefenstette [Gre94]. Deseja-se, aqui, comparar o desempenho das técnicas em sua forma originalmente proposta pelos autores.

Também foram gerados dois tesauros T4 e T5, utilizando o método de Yang e Powers [YP08], sendo o primeiro, T4, criado da forma tradicional, utilizando a medida do cosseno para obter termos similares. O segundo tesauro, T5, foi gerado com base na medida de similaridade de Jaccard, permitindo assim uma comparação com o resultado obtido com a aplicação do método de Grefenstette [Gre94], que utiliza a mesma medida para computar a similaridade dos termos, e permitindo a avaliação da qualidade dos termos relacionados, indiferentemente da métrica de similaridade, e baseada apenas na aplicação de uma técnica adaptada da LSA.

Ao gerar os tesauros, observou-se que existiam alguns termos-chave que continham termos relacionados que ocorriam em mais de um tesauro. Retornou-se então ao conjunto inicial e removeram-se as duplicatas, utilizando-se apenas uma ocorrência do termo para a avaliação. Do total de 456 termos gerados para serem remetidos para análise, restaram 387, após a remoção das duplicatas.

A Tabela 5.1 apresenta a frequência dos termos relacionados gerados para cada um dos métodos apresentados. A princípio, todos os termos deveriam gerar dez termos relacionados para cada termo-chave, porém, como podem ser observados em negrito, os termos “*children*” e “*data_protection*” não geraram a quantidade de termos relacionados pré-estabelecida.

O termo “*children*” gerou uma quantidade menor de termos do que a pré-estabelecida, pois as relações contextuais foram reduzidas quando se realizou o corte para a geração do tesauro, restando apenas seis termos relacionados, como pode ser observado em negrito na linha “*children*”, na Tabela 5.1. Esta baixa quantidade de termos relacionados prejudica a qualidade do tesauro, como veremos na subseção 6.1. Essa quantidade de termos é corrigida através dos métodos de construção de tesauro T3, T4 ou T5, pois o método que gera o tesauro T3 não realiza o corte dos contextos sintáticos,

aumentando assim, o número de relacionamentos entre os termos e obtendo uma melhor qualidade no tesauro.

Tabela 5.1. Quantidade de termos gerados para cada tesauro

Termo	T1	T2	T3	T4	T5	Total
children	10	**6**	10	10	10	46
consent	10	10	10	10	10	50
customer	10	10	10	10	10	50
data_protection	10	**0**	**0**	**0**	**0**	10
data_subject	10	10	10	10	10	50
marketing	10	10	10	10	10	50
notice	10	10	10	10	10	50
personal_data	10	10	10	10	10	50
personal_information	10	10	10	10	10	50
regulation	10	10	10	10	10	50
Total	100	86	90	90	90	456

Os métodos T4 e T5 fazem uso da adaptação da técnica de LSA para a geração dos termos, e com isso, descobrem relações entre os termos que antes não eram aparentes. Com o uso da adaptação da técnica da LSA a quantidade de relações entre termos aumenta, porém isso não significa que a qualidade do tesauro gerado melhore, isto é, embora a adaptação da técnica de LSA descubra novas relações entre termos, isso não significa que estes novos termos encontrados serão similares, conforme será apresentado na Seção 6.

O termo “*data_protection*”, nos tesauros que utilizam a abordagem linguística para encontrar os termos compostos, não foi encontrado. Quando procurado o termo “*data_protection*” no corpus observou-se que o mesmo era sempre modificado por outros termos, como por exemplo, “*personal data protection*” ou “*data protection act*”.

Na extração, foi priorizadoela sempre extrai o termo composto completo, não havendo a ocorrência de “*data_protection*” sem algum modificador no corpus. Porém, utilizando a abordagem de Kaji *et al*. [KMAY00], os termos passam por um processo de desambiguação estrutural, conforme explicado na Seção 3.1.3. Este processo de desambiguação estrutural verificou que o termo “*data_protection*” ocorria mais frequentemente que termos como “*personal_data_protection*” ou “*data_protection_act*”.

Portanto, ao invés de obtermos um termo composto como “*data_protection_act*”, a técnica de desambiguação estrutural obteve o termo “*data_protection*”.

Para este caso, a desambiguação estrutural obteve mais sucesso que uma abordagem linguística. Porém, se necessitássemos encontrar o termo “*personal_data_protection*” utilizando a abordagem com desambiguação estrutural, não encontraríamos, já que ela transformaria o trigrama citado no bigrama “*data_protection*”.

Uma regra intuitiva para solucionar este problema é utilizar ambas as técnicas em conjunto. Dessa forma, quando obtemos um termo composto e, dentro deste termo composto, existe outro termo composto mais frequente, assinalamos o relacionamento de ambos com os outros termos relacionados.

5.3 Sistema de Recuperação de Informações e Sistema de Avaliação

Para a avaliação da similaridade entre os termos-chave e seus termos relacionados, foi solicitado aos especialistas de domínio que julgassem a similaridade dos mesmos de acordo com os contextos em que eles ocorriam. Para isso, o sistema de avaliação foi embutido em um sistema de Recuperação de Informações (RI) desenvolvido no âmbito do projeto com a empresa parceira. Este sistema inclui o corpus utilizado para fazer a geração dos tesauros, uma ontologia de domínio criada manualmente, contendo um total de 112 conceitos, e um glossário de termos relacionados à área de privacidade, contendo um total de 283 conceitos. Parte desta ontologia é apresentada no trabalho de Bruckschen *et al.* [BNS+10].

Nesse sistema de RI o usuário pode navegar entre os documentos, ontologia e termos do tesauro, de forma dinâmica. Com isso, o usuário pode selecionar um termo do tesauro e escolher a opção para localizar o termo selecionado no corpus, na ontologia ou no glossário, caso o termo exista em algum dos mesmos.

A Figura 5.2 apresenta a tela de visualização do tesauro, onde os termos do lado esquerdo são termos-chave e os termos do lado direito são os termos relacionados a cada termo-chave. Na mesma figura, pode-se observar o termo-chave selecionado “*personal_information*” e os termos relacionados “*health_information*”, “*sensitive_information*”, “*patient_records*” etc.

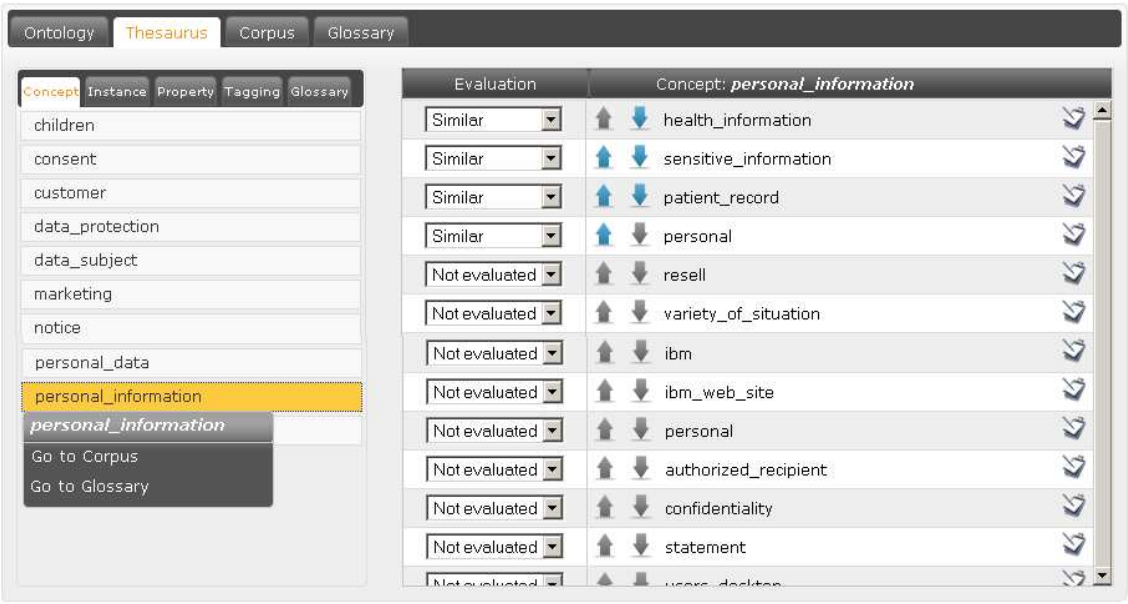

Figura 5.2. Opções para visualização do termo nos recursos

Observando a Figura 5.2 ainda pode-se verificar que o usuário tem as opções de visualizar a região onde o termo "*personal_information*" ocorre no corpus e no glossário. Caso selecionada a opção do corpus, aparecem todas as ocorrências do termo no mesmo, apresentando o nome do documento e a linha em que o termo ocorre, na forma de um concordanceador. A Figura 5.3 apresenta a interface de visualização das ocorrências do termo quando o mesmo é procurado no corpus.

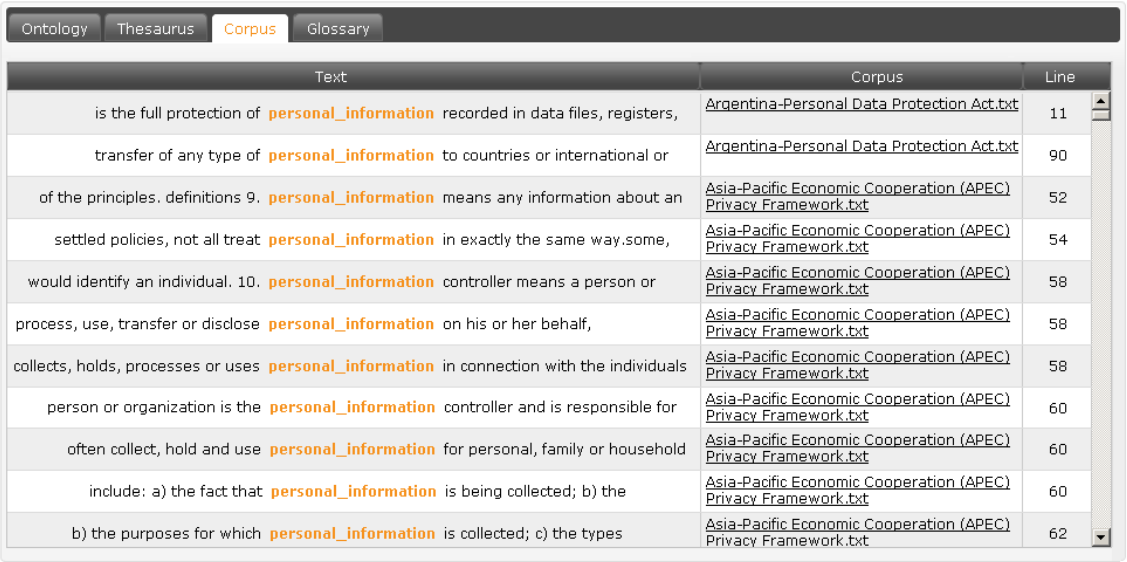

Figura 5.3. Termo "*personal_information*" encontrado no corpus

O mesmo processo de visualização dos termos nos documentos serve para os termos relacionados do tesauro, facilitando assim, ao avaliador, encontrar o contexto em que os termos aparecem, provendo um maior entendimento do termo relacionado.

5.4 Processo de assinalamento do resultado da avaliação

No processo de assinalamento da avaliação, o avaliador deve proceder à seleção do termo-chave e identificação dos termos relacionados gerados como similares, (escolhendo a opção "*Similar*"), ou não similares (escolhendo a opção "*Not similar*") conforme apresentado na Figura 5.4. Após a identificação dos termos relacionados, o avaliador deve classificar os termos similares, indicando se o mesmo é mais ou menos similar, comparado aos outros termos relacionados. Para isso, o avaliador utiliza as setas ao lado do termo relacionado, classificando os termos decrescentemente na ordem de similaridade, de forma que os mais similares fiquem no início da lista.

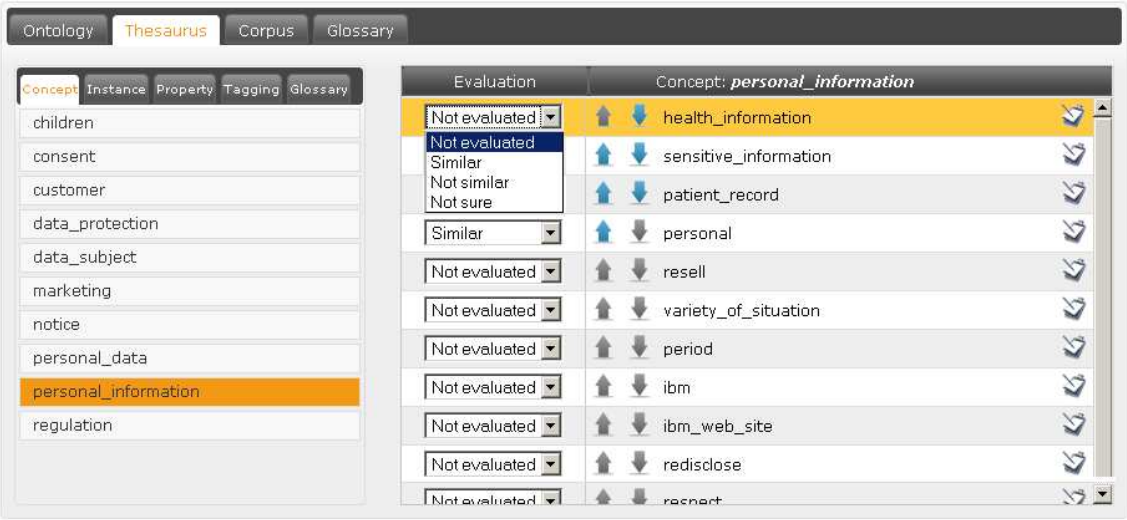

Figura 5.4. Processo de avaliação dos termos relacionados

Caso o usuário não tenha certeza da similaridade de um termo relacionado com o seu termo-chave, pode selecionar a opção "*Not sure*" entre as opções de similaridade. Seja quando o avaliador selecionar esta opção, ou quando desejar fornecer alguma explicação a respeito da classificação de um determinado termo, o mesmo poderá escrever um comentário associado ao termo. Para isso, basta o usuário selecionar o ícone de "escrever comentário", localizada no canto direito de cada termo relacionado, conforme apresentado na Figura 5.5.

Todos os dados da avaliação, isto é, a marcação de similar, não similar ou "*Not sure*", a posição do termo comparado com os outros termos relacionados na lista, comentários e o método pelo qual o tesauro foi gerado, ficam gravados em um arquivo XML. Este arquivo é carregado toda vez que o usuário acessa o sistema, não sendo necessário o usuário avaliar todo o tesauro em uma sessão de utilização do sistema.

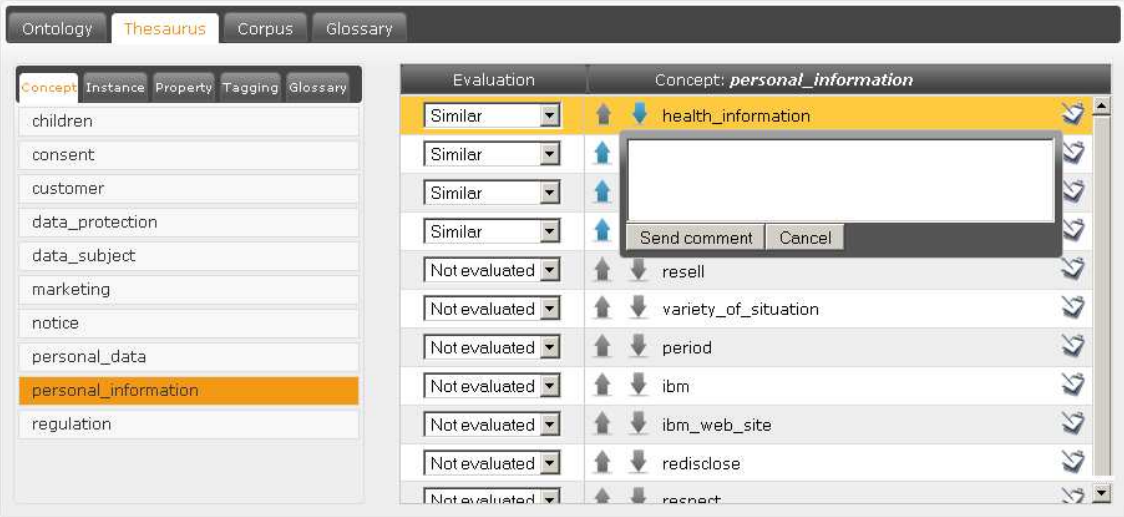

Figura 5.5. Campo para comentário do avaliador

6. RESULTADOS

Os resultados das avaliações dos especialistas de domínio para cada um dos termos gerados por cada um dos métodos são analisados abaixo. Os resultados são analisados em termos das respostas de cada especialista, respostas agrupadas por tesauros e respostas agrupadas pelo conjunto de especialistas. Também são comentadas características referentes ao processo de implementação de cada um dos métodos.

6.1 Resultados por especialista

Os resultados aqui analisados são compilados de forma individual, sendo observadas as avaliações de cada tesauro pela visão do especialista. A primeira análise realizada foi em relação aos termos similares, não similares ou que o avaliador não tem certeza sobre a similaridade, apresentando, além da quantidade de termos, o percentual que esta quantidade representa no total de termos de cada tesauro.

Após, é apresentada uma análise do sentido semântico dos termos relacionados em relação ao tipo de tesauro que foi construído. Este tipo de análise permite verificar quais tesauros geram termos que contém um significado mais próximo ao do termo-chave. Para isso, utilizou-se a classificação dos termos relacionados, realizada por cada especialista, e desta classificação foram gerados gráficos que apresentam a quantidade de termos relacionados em cada tesauro para cada uma das posições da classificação realizada pelo avaliador.

Foram gerados gráficos utilizando as primeiras 25 posições de classificação de cada avaliador, obtendo uma visão ampla do comportamento dos termos para cada tesauro. A seguir, foram gerados gráficos para as primeiras 10 posições de classificação, observando quais tesauros tinham os termos melhor classificados.

- Avaliador 1:

O avaliador 1 foi o único a fazer uso da opção “*Not sure*”, indicando, assim, que não poderia avaliar a real similaridade do mesmo. Isso acontece com termos relacionados em que o significado pode estar associado ao termo-chave dependendo do contexto.

Um exemplo dessa situação ocorre com o termo relacionado “*advertisement*” (em português, “propaganda”), que foi marcado como “*Not sure*” para o termo-chave “*children*” (em português, “crianças”). Inicialmente, poderíamos inferir que o termo “*advertisement*”

não tem relação com o termo “*children*”, pois para o segundo seria esperado encontrar termos relacionados como “*boy*” (em português, “menino”), “*girl*” (em português, “menina”) etc. porém como estamos tratando de documentos do domínio legal e entre os documentos temos leis como “*Children’s online privacy protection rule*”, que trata a respeito da privacidade para crianças, ou como “*Regulation rule pursuant to the telephone disclosure and dispute resolution act of 1992*” que trata de propagandas direcionadas a crianças, conforme pode ser visto no Fragmento 11.

Fragmento 11. Lei que trata de propagandas para menores de 12 anos

> (3) For the purposes of this regulation, if competent and reliable audience composition or readership data does not demonstrate that more than 50% of the audience or readership is composed of children under 12, then the Commission shall consider the following criteria in determining whether an advertisement is directed to children under 12:
>
> (i) Whether the advertisement appears in a publication directed to children under 12, including, but not limited to, books, magazines and comic books;

Dessa forma a associação entre os termos “*advertisement*” e “*children*” não é trivial de ser identificada como similar ou não similar, conforme comentado pelo avaliador. Embora possam existir essas dúvidas, os outros avaliadores preferiram escolher entre “similar” ou “não similar” para termos relacionados, visto que, entre eles, não houve nenhum termo marcado como “*Not sure*”.

O avaliador 1 julgou um total de 387 termos. Destes, 66 termos (17,1%) o avaliador não soube julgar se eram similares ou não similares ao termo-chave. Separando as avaliações por tesauro construído, isto é, contando as repetições, obtemos um total de 456 termos avaliados, e destes, 72 termos (15,8%) marcados como “*Not sure*”.

Devido a os outros avaliadores não terem marcado “*Not sure*” nos tesauros avaliados, serão levados em consideração para a qualidade dos tesauros apenas os termos marcados como similares. Ao final da seção são comentados alguns dados marcados como “*Not sure*” pelo avaliador 1 e as avaliações dadas pelos outros avaliadores. A lista completa de termos julgados pelo avaliador pode ser vista no Apêndice B.

A Tabela 6.1 apresenta a lista completa de percentuais de termos avaliados pelo especialista como “*Similar*”, “*Not similar*” e “*Not sure*”. Esta tabela apresenta as avaliações separadas por tesauro, podendo assim mostrar qual tesauro apresentou mais termos

similares. Com isso podemos ver a eficiência de um método para gerar os termos relacionados, e ainda comparar os métodos utilizados.

Tabela 6.1. Quantidade de termos selecionados pelo avaliador 1 para cada tesauro

	T1	**T2**	**T3**	**T4**	**T5**	**Total**
Similar	44 (44%)	22 (25,6%)	54 (60%)	29 (32,2%)	29 (32,2%)	177(38,8%)
Not similar	36 (36%)	53 (61,7%)	27 (30%)	47 (52,2%)	42 (46,7%)	207 (45,4%)
Not sure	20 (20%)	11 (12,7%)	9 (10%)	14 (15,6%)	19 (21,1%)	72 (15,8%)
Total:	100	86	90	90	90	456

Uma primeira análise nos permite observar que o tesauro que teve o melhor desempenho, isto é, o tesauro que conteve mais termos relacionados avaliados como "*Similar*", foi o tesauro T3, com 60% dos termos marcados como similares. Por outro lado, o tesauro que teve o pior desempenho foi o tesauro T2, apresentando apenas 25,6% de termos marcados como similares. A diferença entre os tesauros T2 e T3 está apenas no corte nos contextos sintáticos. Isso mostra que, para o avaliador 1, fez uma grande diferença o corte dos contextos sintáticos.

Embora este corte não fosse necessário para a computabilidade pelo método de Grefenstette [Gre94], mas apenas para a utilização da adaptação da técnica de LSA, é interessante observar que ele provoca a perda de termos que estariam entre termos similares, antes de computar a matriz através do método de Yang e Powers [YP08].

Outra análise feita é a da eficiência da adaptação da técnica de LSA, aplicada sobre estes termos, pois a quantidade de termos similares passou de 25,6% (tesauro T2) para 32,2% (tesauros T4 e T5). Observa-se que a adaptação da técnica de LSA conseguiu descobrir relacionamentos semânticos que embora não existissem mais devido ao corte, ainda existiam na matriz de contextos.

A eficiência da adaptação da técnica de LSA pode ser observada em termos como, por exemplo, "*person*" encontrado como termo relacionado ao termo-chave "*customer*" no tesauro T3. Devido ao corte dos contextos sintáticos, o termo "*person*" não aparece na lista de termos relacionados do tesauro T2, porém como este termo tinha um significado em outros contextos, a adaptação da técnica de LSA conseguiu encontrar um significado para o mesmo, adicionando este termo ao tesauro T5 como termo relacionado.

Quanto à métrica de similaridade utilizada (tesauros T4 e T5), para o avaliador 1 não pareceu haver diferença entre a métrica do Cosseno e Jaccard. Embora não levemos

em conta a quantidade de termos marcados como "*Not sure*", já que estes termos poderiam mascarar alguma diferença na aplicação da métrica. Para os outros avaliadores fica mais nítida a comparação, visto que não existem termos marcados como "*Not sure*".

Levando em conta apenas os termos marcados como "*Similar*", traçou-se um gráfico para verificar a qualidade dos termos gerados em cada tesauro, isto é, de que tesauro é proveniente a maior quantidade de termos marcados como similares. Este gráfico é apresentado na Figura 6.1, onde no eixo vertical está a quantidade de termos relacionados existentes em cada um dos tesauros. No eixo horizontal está a classificação realizada pelo avaliador.

Um detalhe a ser observado é que o gráfico traz uma representação cumulativa, isto é, são somadas as quantidades de termos conforme aumentam as posições. Assim, para gerar a quantidade de termos similares até a décima posição, o gráfico leva em conta a quantidade de termos similares desde a primeira posição até a décima posição.

Ao analisar, por exemplo, a quantidade de termos relacionados gerados pelo tesauro T1, levando em conta as cinco primeiras posições classificadas pelo avaliador, buscamos o ponto da linha T1, com o número cinco no eixo horizontal. A partir deste ponto, verifica-se no eixo vertical a quantidade de termos relacionados pelo tesauro.

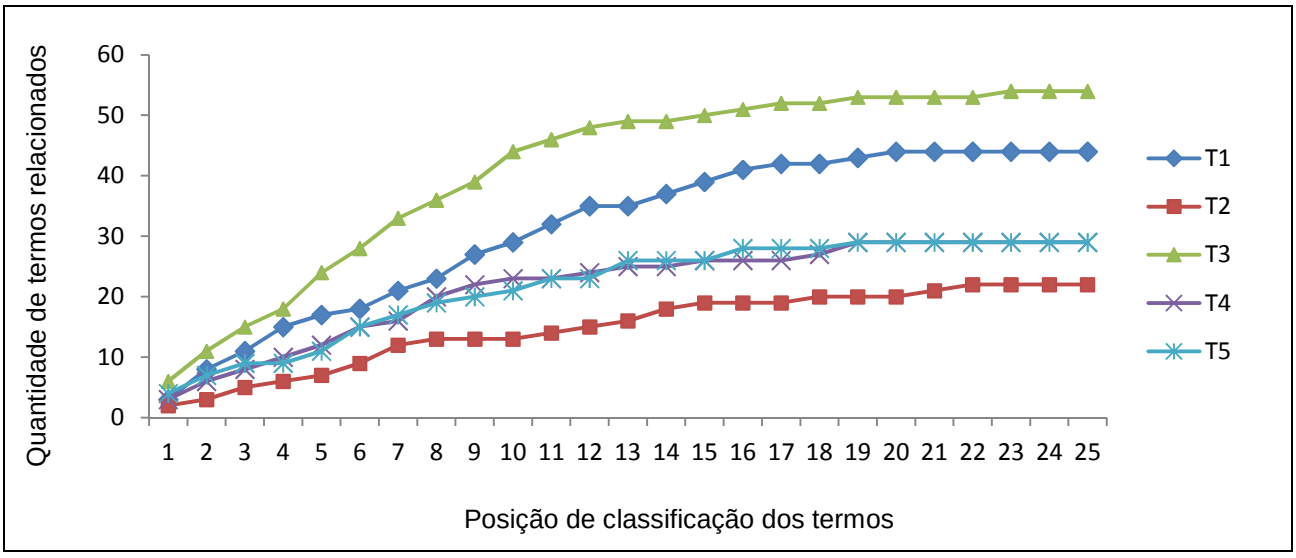

Figura 6.1. Classificação dos termos relacionados segundo o avaliador 1

Analisando as curvas traçadas no gráfico da Figura 6.1, observa-se que o tesauro T3 contém a maior quantidade de termos relacionados em todas as posições. Isso mostra a queda da qualidade dos tesauros que sofreram o corte de contextos. O segundo tesauro que teve a maior quantidade de termos foi o tesauro T1, que é gerado através de método

estatístico, não necessitando de extração de contextos sintáticos. O tesauro que teve a menor quantidade de termos gerados em todas as posições foi o tesauro T2, mostrando que o corte de contextos e a não adaptação da técnica de LSA recuperam poucos termos similares.

Os tesauros que diferem apenas pela métrica de similaridade (T4 e T5) alcançam um desempenho semelhante para os termos, independente da métrica de similaridade utilizada.

Para melhor analisar os termos nas primeiras posições, o gráfico da Figura 6.2 apresenta as dez primeiras posições de cada um dos tesauros. Neste gráfico pode ser observado que o tesauro T3 obtém a maior quantidade de termos nas primeiras dez posições. Assim, observa-se que, dos 54 termos obtidos até a 25ª posição, 44 (81,5% dos termos) se encontram entre as dez primeiras posições.

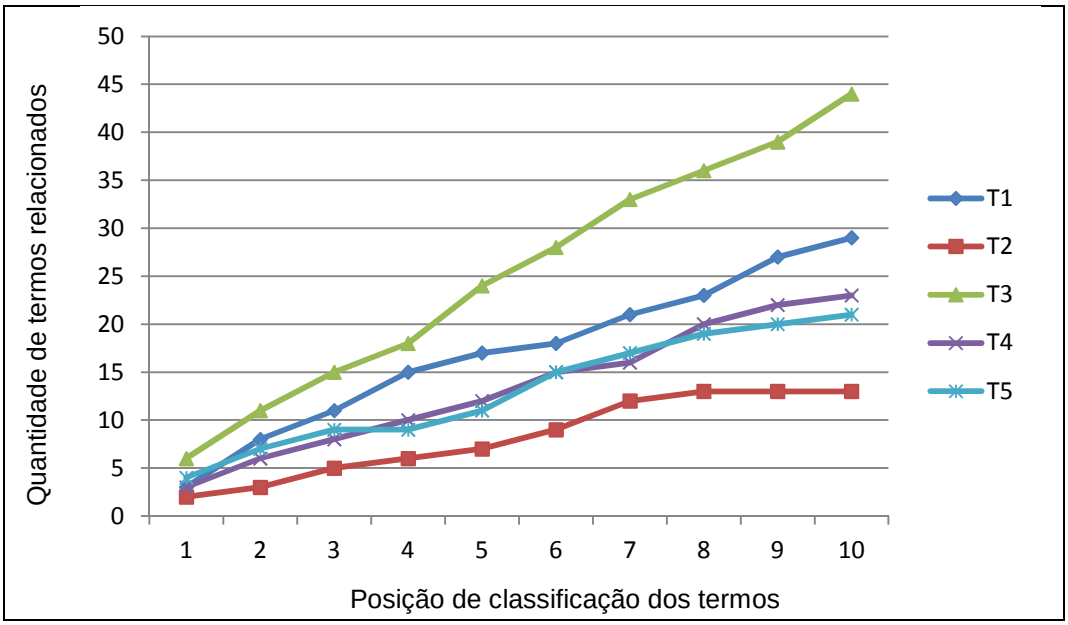

Figura 6.2. Classificação dos 10 primeiros termos relacionados segundo o avaliador 1

Por outro lado, observamos que o tesauro T1, além de conter uma quantidade menor de termos relacionados como similares, gera termos semanticamente menos similares que o tesauro T3, isto é, tem uma curva de crescimento mais suave que a curva gerada pelo tesauro T3. Assim, observa-se que, dos 44 termos contidos até a 25ª posição, 22 (66% dos termos) estão entre os dez primeiros termos.

Ainda seguindo a classificação realizada pelo avaliador 1, podemos concluir que, para um tesauro mais específico, o método de Grefenstette [Gre94] sem a realização do corte de contextos (tesauro T3), poderia ser o mais adequado.

Finalmente, para o avaliador 1, a opção pela métrica de similaridade em um tesauro que utiliza uma adaptação da técnica de LSA não parece promover uma grande diferença na quantidade e na qualidade dos termos gerados.

- Avaliador 2:

O avaliador 2 julgou um total de 387 termos que, quando separados em tesauros, gerou um total de 456 termos avaliados. O avaliador 2 efetuou seu julgamento utilizando apenas as opções "*Similar*" e "*Not similar*", não manifestando, em caso algum, dúvida sobre a similaridade. Partindo destas avaliações, o tesauro que teve a maior quantidade de termos julgados como similares foi o tesauro gerado com o método de Grefenstette [Gre94] (tesauro T3), com um total de 71,1% dos termos marcados como similares, conforme pode ser observado na Tabela 6.2.

Tabela 6.2. Quantidade de termos selecionados pelo avaliador 2 para cada tesauro

	T1	**T2**	**T3**	**T4**	**T5**	**Total**
Similar	62 (62%)	41 (47,7%)	64 (71,1%)	34 (37,8%)	43 (47,8%)	244(53,5%)
Not similar	38 (38%)	45 (52,3%)	26 (28,9%)	56 (62,2%)	47 (52,2%)	212 (46,5%)
Not sure	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Total:	100	86	90	90	90	456

Ainda na Tabela 6.2, pode-se observar que o tesauro que teve a menor quantidade de termos julgados como similares foi o tesauro baseado no trabalho de Yang e Powers [YP08] utilizando a métrica de similaridade do Cosseno (tesauro T4), com um total de 37,8% dos termos marcados como similares.

Fazendo uma comparação entre os tesauros gerados com e sem a adaptação da técnica de LSA, vemos que o tesauro gerado sem a técnica de LSA (tesauro T2) teve um desempenho melhor ou pelo menos comparável às técnicas que utilizam a LSA, fazendo diferença na aplicação da métrica de similaridade.

Comparando o tesauro T2 com o tesauro em que ocorre a aplicação da LSA antes de gerar os termos relacionados e a utilização da métrica de similaridade do Cosseno, o tesauro T2 mostrou um melhor desempenho, com 47,7% dos termos marcados como similares, contra 37,8% dos termos do tesauro T4.

Porém, quando utilizamos a métrica de similaridade de Jaccard na adaptação da técnica de LSA, observamos que a quantidade de termos similares aumenta de 34 termos

(tesauro T4) para 43 termos (tesauro T5), obtendo, dessa forma, um desempenho maior que sem a aplicação da adaptação da técnica de LSA (tesauro T2). Dessa forma, caso fossemos escolher uma métrica de similaridade para a aplicação na geração dos termos semelhantes, segundo a análise dos resultados pelo avaliador 2, seria recomendável a utilização da medida de Jaccard ao invés da aplicação da métrica do Cosseno.

Uma análise mais profunda dos termos reconhecidos como similares pode ser realizada através da observação do gráfico apresentado na Figura 6.3. Este gráfico apresenta a classificação realizada pelo avaliador para as primeiras 25 posições de classificação do avaliador, onde cada curva representa a quantidade de termos relacionados conforme aumenta a posição da classificação.

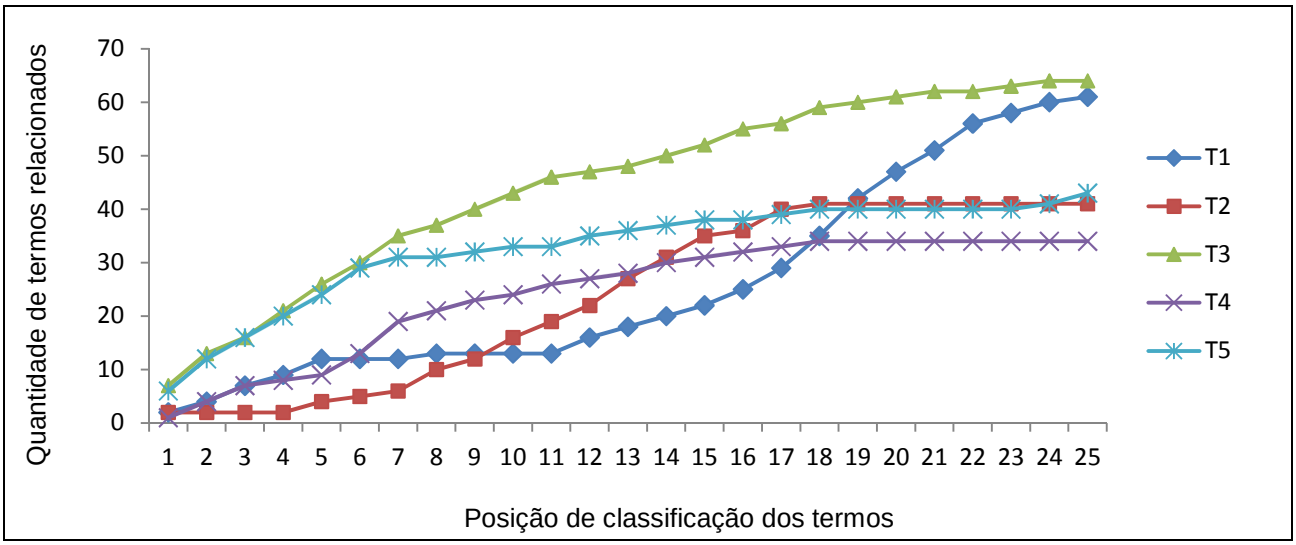

Figura 6.3. Classificação dos termos relacionados segundo o avaliador 2

Como pode ser observado na Figura 6.3, os tesauros T3 e T5 tem a maior quantidade de termos similares nas primeiras posições da classificação. Porém, conforme as posições da classificação aumentam, o tesauro T5 praticamente se estabiliza e o tesauro T3 passa a ter uma curva suavizada quando comparada com as primeiras posições. Enquanto isso, o tesauro T1, que não contêm muitos termos nas primeiras posições, cresce a partir da décima posição.

Podemos observar, também, que o avaliador 2 notou diferença nos resultados da métrica de similaridade utilizada para a geração dos tesauros T4 e T5. De acordo com a classificação realizada pelo avaliador 2, o tesauro que utiliza a métrica de similaridade de Jaccard (tesauro T5) obteve maior quantidade de termos semelhantes nas primeiras 25 posições.

Observando as primeiras dez posições deste gráfico, o que pode ser melhor visto na Figura 6.4, vemos que o tesauro T3, além de conter a maior quantidade de termos marcados como "*Similar*", também contém os termos mais bem classificados nas primeiras posições.

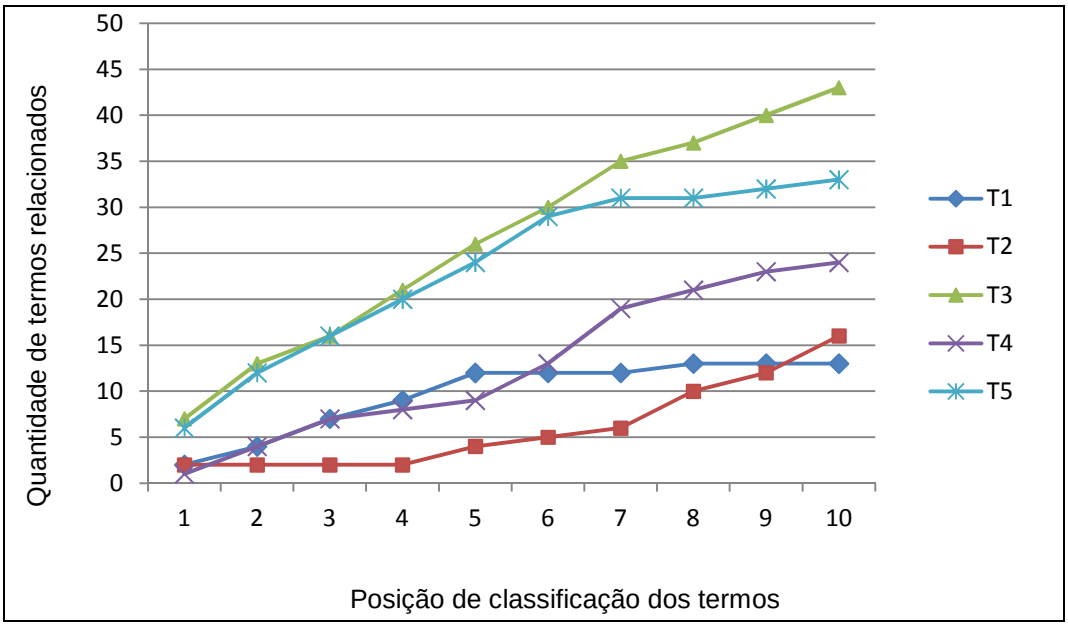

Figura 6.4. Classificação dos 10 primeiros termos relacionados segundo o avaliador 2

O comportamento semelhante, entre os tesauros T3 e T5, nas primeiras posições do gráfico, se deve ao fato de ambos compartilharem termos relacionados. Analisando-se os resultados, observou-se que alguns dos termos que aparecem em ambos os tesauros, foram removidos do tesauro T2 devido ao corte de contextos, mas foram recuperados no tesauro T5 com a adaptação da técnica de LSA e a métrica de similaridade de Jaccard.

Seguindo a classificação realizada pelo avaliador 2, podemos concluir que o método de Grefenstette [Gre94] sem o corte de contextos (tesauro T3) obtém um bom desempenho para a geração de tesauros. Através dessa técnica é possível a geração de uma grande quantidade de termos similares e, ainda, estes termos têm uma forte similaridade semântica com o termo-chave.

- Avaliador 3:

O avaliador 3, assim como o avaliador 2, não assinalou nenhum dos termos avaliados como "*Not Sure*", julgando os termos apenas como similares ou não similares. A Tabela 6.3 apresenta os valores referentes à quantidade de respostas assinaladas pelo avaliador.

Nesta tabela podemos observar que o tesauro que obteve o melhor desempenho foi o tesauro T3, com um total de 70% dos termos assinalados como similares. Em compensação, o tesauro T2 foi o tesauro que obteve o pior desempenho, tendo apenas 33,7% dos termos marcados como similares.

Tabela 6.3. Quantidade de termos selecionados pelo avaliador 3 para cada tesauro

	T1	**T2**	**T3**	**T4**	**T5**	**Total**
Similar	36 (36%)	29 (33,7%)	63 (70%)	41 (45,6%)	42 (46,7%)	211 (46,3%)
Not similar	64 (64%)	57 (66,3%)	27 (30%)	49 (54,4%)	48 (53,3%)	245 (53,7%)
Not sure	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Total:	100	86	90	90	90	456

Comparando com os outros avaliadores, o avaliador 3 tem um perfil de respostas semelhante ao do avaliador 1, apresentando semelhança na avaliação entre os tesauros que têm a maior e menor quantidade de termos semelhantes, bem como a não diferenciação entre métricas de similaridade aplicadas na adaptação da técnica de LSA (tesauros T4 e T5).

Mais uma vez observamos, pelas respostas do avaliador 3, que o tesauro gerado pelo método de Grefenstette [Gre94] com corte nos contextos sintáticos obteve um baixo desempenho quando comparado com o tesauro gerado pelo mesmo método sem o corte nos contextos.

Na ótica desse avaliador observa-se uma queda na classificação do tesauro T1, relativa à quantidade de termos semelhantes quando comparado com os outros avaliadores. Para o avaliador 1, por exemplo, o método que gera o tesauro T1 tinha obtido uma grande quantidade de termos semelhantes, sendo o segundo tesauro com maior quantidade dos mesmos. Para o avaliador 3, o método que gera o tesauro T1 foi o segundo tesauro que gerou a menor quantidade de termos.

Se comparado ao avaliador 2, obtemos uma queda maior ainda na avaliação do tesauro T1, passando de um tesauro com 62 termos similares para um tesauro com 36 termos gerados como similares, enquanto que o tesauro T3 obteve uma queda de 64 termos para 63 termos similares entre os mesmos avaliadores.

Observando a Tabela 6.3 podemos fazer uma comparação do tesauro T2 com o tesauro T5, isto é, diferenciando-se apenas na adaptação da técnica de LSA. Nesta comparação, o tesauro T5 obteve um desempenho melhor (46,7%), contra os 33,7% do

tesauro T2, indicando que a técnica de LSA pode encontrar relações semânticas entre os termos que antes não existiam.

Depois de comparadas as quantidades de termos similares encontrados para cada tesauro, partimos para a análise da classificação dos termos gerados em cada um dos tesauros, verificando se o tesauro, além de gerar uma grande quantidade de termos, também gera termos semanticamente similares.

A Figura 6.5 apresenta um gráfico com a quantidade de termos gerados em cada tesauro, de acordo com as posições que os mesmos ocupam na avaliação. Para essa análise utilizou-se a classificação realizada pelo especialista, que ordenou os termos por ordem decrescente de significado, sendo utilizadas para a criação do gráfico as 25 primeiras posições da classificação realizada pelo avaliador 3.

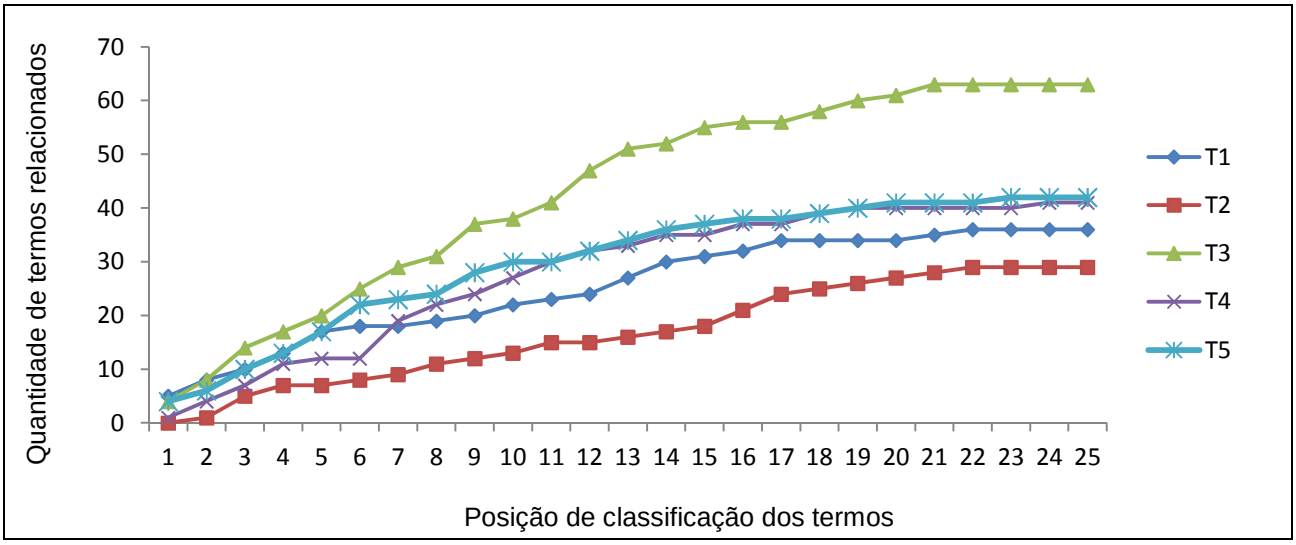

Figura 6.5. Classificação dos termos relacionados segundo o avaliador 3

Analisando esse gráfico, observamos que, para o avaliador 3 assim como para os outros avaliadores, o tesauro T3 contém a maior quantidade de termos significativos. Por outro lado, o tesauro T2, além de conter poucos termos relacionados marcados como similares, também contém termos pouco significativos.

Comparando o tesauro T2 com o tesauro T3, observamos que o corte nos contextos também fez diferença para o avaliador 3, pois acabou retirando termos que eram representativos para o tesauro. Este corte, embora produza economia no processamento computacional dos termos, diminui a quantidade de termos semanticamente similares.

Outra comparação a ser analisada é aquela entre os termos relacionados gerados pelo tesauro T4 e pelo tesauro T5, que se diferenciam apenas pela métrica de

similaridade utilizada. Como pode ser observado, inicialmente o tesauro T5 tem um desempenho melhor que o T4, obtendo uma maior quantidade de termos relacionados, porém essa diferença desaparece a partir da 11ª posição. Embora haja uma diferença na quantidade de termos relacionados, ela não é significativa, principalmente com a melhoria na classificação dada pelos avaliadores.

Para observar melhor os termos gerados nas primeiras posições, o gráfico apresentado na Figura 6.6 mostra as primeiras 10 posições da classificação realizada pelo avaliador 3. Neste gráfico podemos observar que, na primeira posição, o tesauro T1 contém mais termos similares do que o tesauro T3 (5 termos no tesauro T1 e 4 termos no tesauro T3), porém na segunda posição ambos os tesauros contêm 8 termos similares e, após esta posição, o tesauro T3 passa a ter uma quantidade maior de termos similares que o tesauro T1.

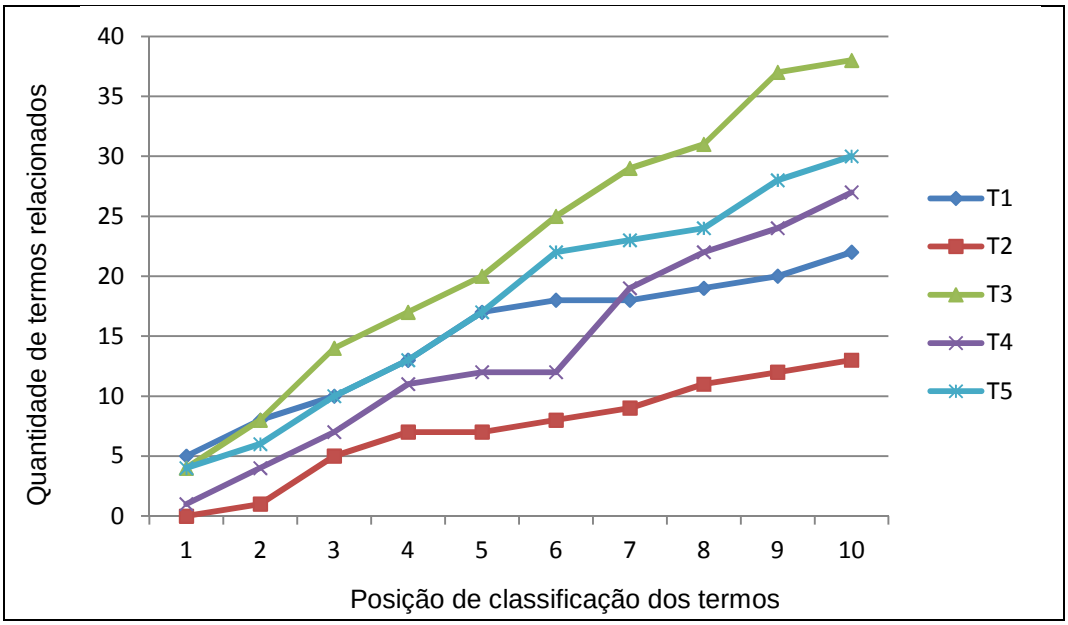

Figura 6.6. Classificação dos 10 primeiros termos relacionados segundo o avaliador 3

Podemos observar que, embora para as 25 primeiras posições (Figura 6.5) a métrica de similaridade não tenha apresentado grande diferença, para as 10 primeiras posições da classificação (Figura 6.6) ela apresenta diferença, obtendo a métrica de Jaccard melhores resultados.

- Comparações dos resultados entre os três avaliadores

Após fazer a análise dos tesauros para cada um dos avaliadores, vamos traçar as principais características observadas em cada uma das avaliações, comparando os resultados obtidos.

A primeira comparação que fazemos é entre os tesauros que obtiveram a maior e a menor quantidade de termos similares. Analisando as tabelas 6.1, 6.2 e 6.3, observamos que o tesauro que teve mais termos julgados como similares foi o tesauro T3. Por outro lado, os tesauros que obtiveram a menor quantidade de termos julgados como similares foram os tesauros T2 e T4. O tesauro T2 teve a menor quantidade de termos julgados como similares por dois dos três avaliadores, mostrando assim que o corte nos contextos teve um grande impacto nos resultados.

O tesauro T4 teve a menor quantidade de termos julgados como similares para um dos avaliadores, mostrando que a métrica de similaridade pode fazer diferença na seleção dos termos para o tesauro. Isso nos leva à segunda comparação, buscando observar qual métrica de similaridade aplicada na adaptação da técnica de LSA seleciona a maior quantidade de termos similares.

Nas tabelas 6.1, 6.2 e 6.3 pode ser observado que a escolha na métrica de similaridade aplicada na construção do tesauro faz diferença na obtenção de termos similares. Comparando a técnica que utiliza a métrica do Cosseno (tesauro T4) e a técnica que utiliza a métrica de Jaccard (tesauro T5), vemos que os resultados de dois dos avaliadores indicam que a métrica de Jaccard tem uma eficiência maior que a métrica do Cosseno. Para um dos avaliadores, a aplicação das métricas não apresentou diferença na quantidade de termos relacionados similares.

Uma última comparação é feita entre os termos gerados através do método de Grefenstette [Gre94] com corte nos contextos (tesauro T2) e o mesmo método porém com a adição da adaptação da técnica de LSA (tesauro T5). Comparando as tabelas 6.1, 6.2 e 6.3, observamos que, em todos os casos, os resultados com a aplicação da adaptação da técnica de LSA obteve melhores resultados, mostrando que a aplicação da adaptação da técnica de LSA pode realmente descobrir relações semânticas que aparentemente não observamos quando olhamos apenas para as frequências dos termos. Com isso, acredita-se que a aplicação da adaptação da técnica de LSA sobre o método de Grefenstette [Gre94] sem corte de contextos, melhoraria ainda mais os resultados.

6.2 Resultados por tesauro

Os resultados apresentados nesta seção são referentes ao tesauro, e não mais à visão do especialista. Para tal análise, optamos pela divisão em dois grupos: união das respostas dos avaliadores, isto é, um termo similar é aquele que foi julgado por um

especialista como “*Similar*”; e interseção, onde um termo é similar caso ele tenha sido avaliado por todos os especialistas como “*Similar*”. A comparação entre essas abordagens visa destacar os tesauros que geram termos de difícil e de fácil identificação, como similares, pelos especialistas. A seguir são apresentados os resultados de cada uma destas abordagens.

- Abordagem da união

Esta abordagem procurou unir todos os termos marcados como similares para cada tesauro, verificando qual tesauro contém a maior quantidade de termos similares. Essa abordagem não consegue identificar o sentido dos termos de cada um dos tesauros (se são termos mais abrangentes ou mais específicos), apenas identificando, dentre todos os termos gerados, quantos são similares.

A ideia desta abordagem é identificar o tesauro que obtém a maior quantidade de termos de uma forma abrangente, pois parte-se do princípio que dois avaliadores discordam da classificação da similaridade de um termo com o termo-chave quando o mesmo é abrangente demais. Dessa forma, termos marcados como “*Similar*” por apenas um dos avaliadores seriam abrangentes, mas ainda contêm algum relacionamento com o termo-chave.

A Tabela 6.4 apresenta a quantidade de termos similares encontrados para cada tesauro. Nessa tabela são apresentadas a quantidade de termos marcados como “S*imilar*” por qualquer um dos avaliadores e o total de termos avaliados para cada tesauro. Podemos observar, nesta tabela, que o tesauro T1 obteve um total de 70% dos termos marcados com “*Similar*”. Este foi seguido pelo tesauro T3, com 64,4% dos termos marcados como similares, indicando que o tesauro T1 tem uma abrangência de termos maior. Por outro lado, o tesauro T5, apresentou 44,4% dos termos marcados como similares.

Tabela 6.4. Quantidade de termos similares na abordagem da união

	T1	**T2**	**T3**	**T4**	**T5**
Similar	70 (70%)	44 (51,2%)	58 (64,4%)	48 (53,3%)	40 (44,4%)

- Abordagem da interseção

Esta abordagem faz a interseção dos resultados indicados pelos avaliadores, selecionando apenas os termos relacionados que aparecem marcados como "*Similar*" em todos os tesauros. Com isso, esta abordagem tenta identificar o tesauro que contém a maior quantidade de termos significativos para o tesauro, partindo da premissa que, se todos os avaliadores marcaram o termo como "*Similar*", o termo realmente tem significado para o tesauro.

A Tabela 6.5 apresenta a quantidade de termos obtidos das interseções das respostas dos avaliadores quando os mesmos julgaram um termo como similar ao termo-chave, para cada tesauro. Como pode ser observado, o tesauro T3 teve a maior quantidade de termos marcados como similar pelos avaliadores, obtendo 51,1% dos termos marcados pelos três avaliadores como similares aos termos-chave. Por outro lado, o tesauro T1 obteve a menor quantidade de termos marcados como similares por todos os especialistas, obtendo um total de 17% dos termos.

Tabela 6.5. Quantidade de termos marcados como "*Similar*" por todos os avaliadores

	T1	**T2**	**T3**	**T4**	**T5**
Similar	17 (17%)	15 (17,4%)	46 (51,1%)	20 (22,2%)	25 (27,8%)

- União & Interseção

Comparando essas duas abordagens, podemos verificar que o tesauro T3 obteve a menor queda na quantidade de termos marcados como similares. De todos os termos marcados como similares pelos avaliadores, apenas 13,3% geraram dúvidas.

Em contrapartida, o tesauro T1 foi o que gerou a maior quantidade de termos em que os avaliadores tiveram dúvidas quanto à similaridade: 53% dos termos gerados provocaram este tipo de dúvidas.

Aqui podemos verificar, ainda, que a métrica de similaridade também apresentou diferença na quantidade de termos marcados como similares. O tesauro T5 obteve uma queda menor na quantidade de termos (16,6%), quando comparado ao tesauro T4 (31,1%). Essa diferença indica que a utilização da métrica de Jaccard como forma de calcular a similaridade entre os termos, gerou termos que tem uma maior chance de serem marcados por todos os especialistas como similares.

6.3 Análises de caso

Nesta seção vamos dar atenção especial para alguns casos particulares, pois acreditamos serem casos interessantes para o entendimento dos métodos. O primeiro caso é o do termo “*children*”, devido a este apresentar uma quantidade menor de termos quando realizado o corte de contextos no método de Grefenstette [Gre94]. Logo a seguir, é analisado o caso das respostas marcadas como “*Not sure*”, visto que apenas um dos avaliadores usou desta alternativa, tentando descobrir se ocorreriam diferenças na avaliação caso o avaliador tivesse marcado outra alternativa.

- Caso “*children*”

A ideia, neste caso, é observar se a não realização de um corte nos contextos sintáticos ou a aplicação da adaptação da técnica de LSA nos termos com a realização do corte melhoram os resultados do tesauro gerado. Para isto, comparamos a quantidade de termos gerados como similares entre os tesauros (excluindo o tesauro T1 por não utilizar nenhuma das técnicas envolvidas).

A Tabela 6.6 apresenta, para o termo-chave “*children*”, a quantidade de termos marcados como “*Similar*” para termos similares, “*Not similar*” para termos não similares e “*Not sure*” para termos que o usuário não soube avaliar.

Tabela 6.6. Avaliações para termos relacionados ao termo-chave "*children*"

		T2	**T3**	**T4**	**T5**
Avaliador 1	***Similar***	1	3	1	0
	Not similar	5	7	9	9
	Not sure	0	0	0	1
Avaliador 2	***Similar***	1	4	0	1
	Not similar	5	6	10	9
	Not sure	0	0	0	0
Avaliador 3	***Similar***	1	5	2	1
	Not similar	5	5	8	9
	Not sure	0	0	0	0

Analisando a Tabela 6.6, podemos observar que, embora a quantidade de termos tenha aumentado com a adaptação da técnica de LSA (tesauros T4 e T5), a qualidade dos termos não apresentou uma melhora significativa.

Somente o tesauro T4, entre os que utilizam a adaptação da técnica de LSA, apresentou um aumento na quantidade de termos similares, porém apenas pelo julgamento do avaliador 3. Embora tenha aumentado em um termo, ele passou a ter outros 3 termos não similares, não valendo a pena a aplicação dessa técnica para o aumento de termos.

Por outro lado, a não realização do corte nos contextos fez aumentar consideravelmente a qualidade dos termos, como pode ser visto no tesauro T3, pois passou de 1 termo similar gerado no tesauro T2 para 5 termos similares no tesauro T3, segundo o avaliador 3, mantendo a mesma quantidade de termos não similares.

Com isso, podemos concluir que, ao invés de utilizar um método usa a adaptação da técnica de LSA para gerar termos relacionados, é melhor aplicar o método de Grefenstette [Gre94] sem a realização do corte de contextos.

- Caso "*Not sure*"

A ideia aqui é observar a utilização dos valores de "*Not sure*" como valores similares ou não similares de acordo com as respostas dos outros especialistas. Dessa forma, caso ambos os outros dois especialistas tenham respondido "*Similar*" para um termo, ele passa de "*Not sure*" para "*Similar*", e caso ambos tenham respondido "*Not similar*", o termo "*Not sure*" é confirmado como "*Not similar*". Com isso podemos verificar se a mudança, na resposta do avaliador 1, iria acarretar diferenças nos resultados dos tesauros gerados.

Tabela 6.7. Identificação dos termos marcados como "*Not sure*" pelo avaliador 1

	T1	**T2**	**T3**	**T4**	**T5**	**TOTAL**
Similar	3	2	3	5	3	17
Not similar	4	4	2	6	7	23
Discordância	13	5	4	3	9	33
TOTAL:	20	11	9	14	19	73

A Tabela 6.7 apresenta a quantidade de termos que foram marcados como "*Not sure*" pelo avaliador 1 e foram marcados pelos outros avaliadores como "*Similar*" ou "*Not similar*". A tabela ainda apresenta a linha "Discordância", que indica a quantidade de

termos que foram marcados por um dos avaliadores como “*Similar*” e pelo outro avaliador como “*Not similar*”, havendo discordância entre os mesmos. Exemplos de termos marcados como “*Not sure*” podem ser vistos no Apêndice B.

Observando a Tabela 6.7, vemos que a maioria dos termos marcados como “*Not sure*” pelo avaliador 1 foram avaliados diferentemente pelos outros avaliadores, havendo discordância entre eles. Ainda assim, podemos verificar se os termos antes marcados como “*Not sure*”, modificam os resultados da comparação dos tesauros para o avaliador 1. Para isso, remontamos a Tabela 6.1, porém ao invés de colocarmos os valores para “*Not sure*”, adicionamos os valores da Tabela 6.7, modificando o valor de “*Not sure*” por valor “Desconhecido” quando os outros avaliadores discordavam da resposta. Os novos valores para os termos marcados como similares e não similares podem ser vistos na Tabela 6.8.

Tabela 6.8. Novos valores de similaridade para o avaliador 1

	T1	**T2**	**T3**	**T4**	**T5**	**Total**
Similar	47 (47%)	24 (27,9%)	57 (63,4%)	34 (37,7%)	32 (35,5%)	194 (42,5%)
Not similar	40 (40%)	57 (66,3%)	29 (32,2%)	53 (58,9%)	49 (54,4%)	228 (50%)
Desconhecido	13 (13%)	5 (5,8%)	4 (4,4%)	3 (3,4%)	9 (21,1%)	34 (7,5%)
Total:	100	86	90	90	90	456

Comparando os valores obtidos na Tabela 6.8, observa-se que pouca coisa mudou com relação aos valores da Tabela 6.1, continuando o tesauro T3 com a maior quantidade de termos selecionados como similares (obtendo um aumento de 3,4%), e o tesauro T2 com a menor quantidade de termos selecionados como similares (obtendo um aumento de 2,3%).

Também podemos observar, comparando as duas tabelas, que houve uma melhora dos resultados obtidos pelo tesauro T4 se comparado com o tesauro T5, que só utiliza a métrica de similaridade como diferença. Neste caso, o tesauro com a métrica de similaridade do Cosseno obteve um melhor desempenho se comparado com a medida de Jaccard.

6.4 Análises de implementação

Nesta seção tecemos alguns comentários com relação à implementação dos métodos do ponto de vista da utilização das ferramentas, dificuldades encontradas e tempos de processamento.

- Ferramentas e implementação

O mais simples e o mais fácil de implementar. Assim podemos definir a implementação do método de Kaji *et al.* [KMAY00] que utiliza apenas técnicas estatísticas para a construção do tesauro. Por utilizar apenas técnicas estatísticas, a criação deste tipo de tesauro necessitou apenas da instalação da ferramenta NSP, que computa desde a extração dos termos, até a medida de Informação Mútua. Para a aplicação desse método foram criadas funcionalidades apenas para o processo de Desambiguação Estrutural, conforme descrito na seção 4.2.1.

O método de Grefenstette [Gre94], assim como o método de Yang e Powers [YP08], utilizam a informação sintática dos termos do corpus. Para isso é necessário fazer a identificação das classes gramaticais e da estrutura sintática no corpus. A ferramenta utilizada para este processo foi o analisador sintático desenvolvido por Stanford. Após a ferramenta fazer a identificação das classes gramaticais e da estrutura sintática dos termos, foi necessário criar uma funcionalidade para a extração das identificações de cada termo.

Para o cálculo da similaridade entre os termos foi utilizada a ferramenta Lingua Toolkit, que permite o cálculo da similaridade com onze métricas diferentes. Para a utilização dessa ferramenta foram criadas duas funcionalidades, a primeira delas para fazer a formatação dos dados de entrada da ferramenta e a segunda para a extração dos termos com as respectivas métricas de similaridade.

Por fim, o método de Yang e Powers [YP08] além de utilizar as ferramentas que o trabalho de Grefenstette [Gre94] utiliza, ainda faz a utilização de uma ferramenta para a Decomposição em Valores Singulares (SVD). Para a SVD foi utilizada a ferramenta matemática Octave e, com isso, mais duas funcionalidades foram criadas. A primeira delas para a criação das matrizes esparsas que serviram de entrada para a ferramenta. A outra foi criada para extrair os valores gerados pela SVD para cada contexto sintático existente.

Este método também foi o que apresentou maior dificuldade na implementação, pois foi necessária a realização de um corte nos contextos sintáticos para o processamento das matrizes, não sendo possível o processamento sem o corte.

Por fim, vemos que conforme aumentamos a quantidade de informações que desejamos extrair do corpus, mais ferramentas são necessárias e, mais difícil e suscetível ao erro o processo se torna.

- Tempos de processamento

Para a verificação dos tempos de cada um dos processos, foi utilizado um computador Pentium 4, CPU 3.20 GHz, contendo 1.49 GB de memória RAM e sistema operacional Linux Ubuntu 9.04. Na Tabela 6.9 são apresentados os tempos para a geração dos tesauros em cada um dos processos. Os tempos de T2 até T5 não levam em conta o tempo de análise sintática do corpus.

Tabela 6.9. Tempos de geração de cada um dos tesauros

Tesauro	**Tempo**
T1	58 min 45 seg
T2	2 min 25 seg
T3	2 min 24 seg
T4	24 min 23 seg
T5	24 min 23 seg

Observamos que os tesauros T4 e T5 obtém o mesmo tempo de processamento. Isto se deve ao fato de diferenciarem-se apenas no fim do processo, ao fazer a extração dos termos relacionados, para a geração do tesauro.

Como os tempos da Tabela 6.9 não incluem os tempos de análise sintática do corpus podemos adicionar o tempo de aproximadamente 29 horas e 30 minutos para o processamento do corpus. O corpus utilizado está separado em cem documentos, contendo um total de 1.122.836 palavras.

Com isso, o processamento para a geração do tesauro T1 passa a ser o que leva menos tempo para ser gerado. O tempo da geração do tesauro T1 ainda pode variar de acordo com o tamanho da janela utilizada. O tempo utilizado apenas para a criação dos termos relacionados, utilizando uma janela de 30 termos, foi de 49 minutos e 6 segundos.

Para a criação de um tesauro que utiliza a anotação sintática do corpus, o tempo é muito maior devido ao tempo de anotação do corpus. Porém, como a anotação sintática pode ser realizada apenas uma vez para todos os tesauros, vemos que os tesauros gerados pelo método de Grefenstette [Gre94] utilizam menos tempo para serem criados.

Levando em consideração o tempo de processamento para a escolha na aplicação de um método de construção automática de tesauro, um cuidado a ser tomado é o processamento do corpus por um analisador sintático, pois observamos que este é o processo que despende maior tempo.

7. CONCLUSÃO

Na conclusão do presente trabalho, trazemos nossas considerações e percepções acerca do trabalho apresentado nesta dissertação, e de seus resultados. Além disso, relacionamos as contribuições científicas deste trabalho e propostas de trabalhos futuros.

7.1 Considerações

Este trabalho estudou processos de construção automática de tesauro, visando à descoberta de características que identifiquem o melhor método de construção para um determinado contexto. Com isso, um usuário pode definir o método que irá utilizar em seu sistema de acordo com as características desejadas, como ênfase na quantidade de termos gerados, ênfase na similaridade dos termos gerados, tempo de processamento etc.

Foram analisadas detalhadamente as respostas de cada um dos avaliadores identificando os métodos que retornavam a maior e a menor quantidade de termos semelhantes. Também foram realizadas análises sobre a classificação dos termos realizada pelos avaliadores, descobrindo de quais tesauros os termos mais bem classificados eram provenientes, e identificando, dessa forma, características de cada um dos métodos utilizados para a construção dos tesauros.

Após, procedeu-se à análise, não mais pelos resultados dos avaliadores individualmente, mas reunindo as respostas dos mesmos. Para isso, utilizaram-se duas abordagens: uma considerando que, caso o termo fosse marcado como similar por algum dos avaliadores, este era considerado similar ao termo-chave; a outra, considerando que um termo só seria similar ao termo-chave caso ele fosse marcado por todos os avaliadores como similar. A comparação dessas abordagens mostra que, muitas vezes, um tesauro pode gerar uma grande quantidade de termos similares, porém difíceis de avaliar quanto a sua similaridade com o termo-chave.

Por fim, fez-se a análise de dois casos que se acreditou serem interessantes. O primeiro deles foi do termo “*children*” que, devido ao corte nos contextos, teve a quantidade de termos relacionados diminuída. Procurou-se verificar a melhor opção entre adicionar a adaptação da técnica de LSA aos termos depois de realizado o corte de contextos, identificando relações semânticas que os termos com o corte não continham,

ou utilizar o método de Grefenstette [Gre94] sem a redução de contextos. Esta última se mostrou com melhores resultados do que a aplicação da adaptação da técnica de LSA.

O segundo caso analisado foi das respostas dadas pelo avaliador 1 como "*Not sure*", utilizando-se do conhecimento dos outros especialistas para aumentar a quantidade de termos similares e não similares nesses casos, observando modificações nos tesauros gerados. Embora a utilização do conhecimento dos outros especialistas tenha aumentado a quantidade de termos similares para os tesauros, esse aumento não provocou modificação no resultado final, mantendo o tesauro T3 com a maior quantidade de termos gerados como similares.

A avaliação qualitativa realizada por especialistas do domínio de privacidade e utilizada neste trabalho mostrou-se extremamente proveitosa, principalmente por contribuir com a análise de termos que são de difícil identificação como semanticamente similares. Sabe-se que, em se tratando de semântica, as respostas são muito subjetivas, e essa subjetividade nos permitiu descobrir o sentido dos termos gerados pelos métodos estudados.

Os resultados mostraram que a adaptação da técnica de LSA apresenta uma melhora nos resultados, se comparados com os dados originais, quando ambos utilizam um corte nos contextos. Por outro lado, é melhor utilizar a técnica de Grefenstette [Gre94] sem o corte nos contextos do que utilizar a adaptação da técnica de LSA com um corte nos mesmos. Ainda, a escolha da métrica de similaridade empregada na adaptação da técnica de LSA se torna importante, mostrando-se a métrica de Jaccard melhor do que a aplicação da métrica do Cosseno no estudo realizado.

Por fim, a análise buscada com este trabalho e as aplicações desenvolvidas foi a de encontrar o melhor método de construção automática de tesauro, avaliada nesse trabalho para o domínio legal. Para outros domínios, seria interessante realizar novos experimentos, conforme proposto na seção de trabalhos futuros.

7.2 Contribuições

Nesta seção, relacionamos algumas das contribuições deste trabalho nos contextos acadêmico e industrial para o conhecimento produzido. São elas:

- Contribuições principais
  - Processos de construção automática de tesauros baseados em um corpus do domínio em questão;

  - Sistemas para construção de tesauros baseada em métodos estatísticos, baseada em métodos que utilizam conhecimento sintático, e baseada em métodos com uso da adaptação da técnica de LSA;
  - Avaliação qualitativa dos resultados obtidos na experimentação do sistema. Esta avaliação sendo realizada com o apoio de especialistas do domínio de privacidade, permitindo uma visão subjetiva dos termos.

- Recursos
  - Corpus *Privacy*, desenvolvido em conjunto com a equipe, no âmbito do projeto APAO, embora não existam quaisquer restrições para a sua utilização em outros projetos e pesquisas;
  - Corpus *Privacy* anotado sintaticamente através do parser desenvolvido em Stanford.

- Artigo
  - “Comparação de técnicas para a construção de tesauros visando o enriquecimento de uma ontologia do domínio legal”, aceito no “3º Seminário de Pesquisa em Ontologias no Brasil – 3º ONTOBRAS”, com resultados preliminares do trabalho até o primeiro semestre de 2010 [GBS+10].

7.3 Trabalhos futuros

No decorrer deste trabalho, algumas ideias de trabalhos futuros baseados neste, foram elaboradas. Algumas destas ideias são detalhadas nesta seção. São elas:

- Geração de tesauro baseada no método de Yang e Powers [YP08] sem o corte nos contextos

  Ao analisar os resultados obtidos, notou-se que a adaptação da técnica de LSA melhorou os resultados dos tesauros gerados se comparados aos mesmos sem a utilização da técnica. Porém, devido a limitações de *hardware* as matrizes antes da execução da LSA tiveram que ser reduzidas, sendo realizado o corte de contextos. Como apresentado nos resultados, esse corte prejudicou as relações entre os termos. Com isso, acreditamos que a aplicação da adaptação da técnica

de LSA sobre os termos, sem corte, melhoraria os resultados obtidos quando comparado com a técnica de Grefenstette [GRe94] (tesauro T3).

- Experimentação dos métodos em um domínio diferente
  Os métodos propostos podem ser aplicados em outros domínios. A aplicação deste trabalho considerou o domínio de privacidade de dados na indústria de software, o que particulariza a avaliação por especialistas de domínio. A aplicação em outro domínio permite verificar se as características de cada tesauro gerado permanecem inalteradas.

- Realizar a construção de uma taxonomia dos termos relacionados, com relação ao termo-chave
  Parte-se do princípio que os termos relacionados são gerados por um tesauro associativo, portanto temos termos relacionados associados semanticamente ao termo-chave, porém isso não nos diz muito sobre o termo-chave. Acredita-se que o refinamento do significado do termo-chave pode ser feito através da criação de uma taxonomia dos termos relacionados, identificando nos mesmos relações melhor definidas como sinonímia, antonímia, meronímia, hiperonímia etc. Essa taxonomia poderia ser utilizada em um sistema de RI, permitindo ao usuário a escolha de recuperar documentos que além de conter o termo procurado, também documentos que contêm merônimos semanticamente relacionados ao termo-chave, por exemplo.

- Experimentação dos métodos utilizando um corpus em outro idioma
  A realização de experimentos utilizando um corpus em outro idioma permitiria verificar se o comportamento dos métodos permanece o mesmo quando o idioma é trocado. Para alguns idiomas adaptações seriam necessárias, como o caso do português em que o contexto sintático de substantivo que modifica substantivo não seria utilizado.

REFERÊNCIAS BIBLIOGRÁFICAS

[AGB02] J. Aitchison, A. Gilchrist, D. Bawden. “Thesaurus construction and use: a practical manual”. Routledge, 2002, 4 ed, 230p.

[AMS08] V.M.P. Anick, V. Murthi, S. Sebastian. “Similar term discovery using web search”. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008, pp. 1209-1213.

[BDO95] M.W. Berry, S.T. Dumais, G.W. O'Brien. “Using Linear Algebra for Intelligent Information Retrieval”. *SIAM Review*, vol. 37-4, Dezembro 1995, pp. 573-595.

[Bin08] H. Bing-Geng. “The Architecture and Algorithms of Retrieval Thesaurus on Web”. In: Proceedings of the International Conference on Computer Science and Software Engineering, 2008, pp. 448-450.

[BNS+10] M. Bruckschen, C. Northfleet, D.M. Silva, P. Bridi, R.L. Granada, R. Vieira, P. Rao, T. Sander. "Named entity recognition in the legal domain for ontology population". In: SPLeT 2010: The 3rd Workshop on Semantic Processing of Legal Texts, 2010, pp. 16-21.

[BP03] S. Banerjee, T. Pedersen. "The Design, Implementation, and use of the Ngram Statistics Package". In: Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, 2003, pp. 370-381.

[CC07] L. Chen, S. Chen. “A New Approach for Automatic Thesaurus Construction and Query Expansion for Document Retrieval”. *International Journal of Information and Management Sciences*, vol. 18-4, Dezembro 2007, pp. 299-315**.**

[CG06] M.L.A. Campos, H.E. Gomes. “Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador”. *Perspectivas em ciência da informação*, vol. 11-3, Set-Dez 2006, pp. 348-359.

[CH90] K.W. Church, P. Hanks. “Word association norms, mutual information, and lexicography”. *Computational Linguistics*, vol. 16-1, Março 1990, pp. 22-29.

[Coo69] T. Cooper. “Thesaurus linguae Romanae et Britannicae, 1565”. Scolar P., 1969, 2000p.

[Cro88] C.J. Crouch. “A cluster-based approach to thesaurus construction”. In: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, 1988, pp. 309-320.

[CY92] C.J. Crouch, B. Yang. "Experiments in automatic statistical thesaurus construction". In: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 1992, pp. 77–88.

[DFL+88] S.T. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, R. Harshman. "Using latent semantic analysis to improve access to textual information". In: Proceedings of the SIGCHI conference on Human factors in computing systems, 1988, pp. 281-285.

[Dum93] S.T. Dumais. “LSI meets TREC: A status report”. In: Proceedings of First Text Retrieval Conference (TREC-1), 1993, pp. 137-152.

[Dum94] S.T. Dumais. "Latent Semantic Indexing (LSI) and TREC-2". In: Proceedings of the Second Text REtrieval Conference (TREC2), 1994, pp. 105-115.

[Dum95] S.T. Dumais. "Using LSI for information filtering: TREC-3 experiments". In: Proceedings of the Third Text REtrieval Conference (TREC3), 1995, pp. 219-230.

[FAT98] K. Frantzi, S. Ananiadou, J. Tsujii. "The C-value/NC-value Method of Automatic Recognition for Multi-word Terms". In: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, 1998, pp. 585-604.

[Fel98] C. Felbaum. “Wordnet, an Electronic Lexical Database”. Cambridge: MIT Press, 1998, 445p.

[Fir57] J. Firth. "A Synopsis of Linguistic Theory 1930-1955". Studies in Linguistic Analysis, 1957, 205p.

[Gas01] C.V. Gasperin. “Extração automática de relações semânticas a partir de relações sintáticas”. Dissertação de mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2001, 137p.

[GBS+10] R.L. Granada, M. Bruckschen, V.L.S. de Lima, R. Vieira, C. Northfleet. "Comparação de técnicas para a construção de tesauros visando o enriquecimento de uma ontologia do domínio legal". In: 3º Seminário de Pesquisa em Ontologias no Brasil, 2010.

[GK65] G. Golub, W. Kahan. "Calculating the singular values and pseudo-inverse of a matrix". *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, vol. 2-2, 1965, pp. 205-224.

[GL03] C.V. Gasperin, V.L.S. de Lima. “Experiments on extracting semantic relations from syntactic relations”. In: Proceedings of the 4th international conference on Computational linguistics and intelligent text processing, 2003, pp. 314-324.

[Gre94] G. Grefenstette. “Explorations in automatic thesaurus discovery”. Kluwer Academic Publishers Norwell, 1994, 306p**.**

[Har54] Z.S. Harris."Distributional structure". *Words*, vol. 10-23, 1954, pp. 146-162.

[HE07] M. Heilman, M. Eskenazi. “Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions”. In: Proceedings of the SLaTE Workshop on Speech and Language Technology in Education, 2007, pp. 65-68.

[Hea92] M.A. Hearst. "Automatic acquisition of hyponyms from large text corpora". In: Proceedings of the 14th conference on Computational Linguistics, 1992, pp. 539-545**.**

[INHN08] M. Ito, K. Nakayama, T. Hara, S. Nishio. “Association thesaurus construction methods based on link cooccurrence analysis for wikipedia”. In: Proceedings of the 17th ACM Conference on Information and Knowledge management, 2008, pp. 817-826**.**

[JC94] Y. Jing, W.B. Croft. "An association thesaurus for information retrieval". In: Proceedings of Recherche d'Information Assistee par Ordinateur - RIAO, 1994, pp. 146-161.

[KHT08] A. Kongthon, C. Haruechaiyasak, S. Thaiprayoon. “Constructing term thesaurus using text association rule mining”. In: Proceedings of 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008, pp. 137-140.

[KM03] D. Klein, C.D. Manning. "Accurate unlexicalized parsing". In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2003, pp. 423-430.

[KMAY00] H. Kaji, Y. Morimoto, T. Aizono, N. Yamasaki. “Corpus dependent association thesauri for information retrieval”. In: Proceedings of the 18th Conference on Computational Linguistics, 2000, pp. 404-410.

[Kna00] S.D. Knapp. "The contemporary thesaurus of social science terms and synonyms: A guide for natural language computer searching". Greenwood, 2000, 2 ed, 656p.

[KY00] A. Kilgarriff, C. Yallop. "What's in a thesaurus". In: Proceedings of the Second International Conference on Language Resources and Evaluation, 2000, pp. 1371-1379.

[LD97] T.K. Landauer, S.T. Dumais. “A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge”. *Psychological review*, vol. 104-2, Abril 1997, pp. 211-240.

[LFL98] T.K. Landauer, P.W. Foltz, D. Laham. “An introduction to Latent Semantic Analysis”. *Discourse processes*, vol. 25-2&3, 1998, pp. 259-284.

[Lin98] D. Lin. "Automatic retrieval and clustering of similar words". In: Proceedings of the 17th international conference on Computational linguistics. 1998, pp. 768-774**.**

[LSP09] C.H. Li, W. Song, S.C. Park. “An automatically constructed thesaurus for neural network based document categorization". *Expert Systems with Applications*, vol. 36-8, Outubro 2009, pp. 10969-10975.

[MH03] D. Mollá, B. Hutchinson. "Intrinsic versus extrinsic evaluations of parsing systems". In: Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?, 2003, pp. 43-50.

[MKM+94] M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, B. Schasberger. "The Penn Treebank: annotating predicate argument structure". In: Proceedings of the workshop on Human Language Technology, 1994, pp. 114-119.

[MMM06] M. de Marneffe, B. MacCartney, C.D. Manning. "Generating Typed Dependency Parses from Phrase Structure Trees". In: Proceedings of the fifth international conference on Language Resources and Evaluation – LREC’06, 2006, pp. 449–454.

[MRS08] C.D. Manning, P. Raghavan, H. Schutze. “An introduction to Information Retrieval”. Cambridge University Press, 2008, 496p.

[MS99] C.D. Manning, H. Schutze. “Foundations of statistical natural language processing". MIT Press, 1999, 680p.

[Mil95] G.A. Miller. "WordNet: a lexical database for English". *Communications of the ACM*, vol. 38-11, Novembro 1995, pp. 39-41**.**

[ML08] S.M.W. Moraes, V.L.S. de Lima. "Abordagem não supervisionada para extração de conceitos a partir de textos". In: Proceedings of the XIV Brazilian Symposium on Multimedia and the Web - WebMedia'08, 2008, pp. 359-363.

[NHN07] K. Nakayama, T. Hara, S. Nishio. “Wikipedia mining for an association web thesaurus construction”. In: Proceedings of the 8th international conference on Web information systems engineering, 2007, pp. 322-334.

[NIHN08] K. Nakayama, M. Ito, T. Hara, S. Nishio. "Wikipedia Mining for Huge Scale Japanese Association Thesaurus Construction". In: Proceedings of the 22nd International Conference on Advanced Information Networking and Applications, 2008, pp. 1150-1155.

[PL03] L.A.S. Pizzato, V.L.S de Lima. "Evaluation of a thesaurus-based query expansion technique". In: Proceedings of the 6th international conference on Computational processing of the Portuguese language - PROPOR'03, 2003, pp. 251-258.

[PS07] S.P. Ponzetto, M. Strube. "Deriving a large scale taxonomy from Wikipedia". In: Proceedings of the national conference on Artificial Intelligence, 2007, pp. 1440-1445.

[QGS06] A. Quarteroni, P. Gervasio, F. Saleri. "Scientific computing with MATLAB and Octave". Springer, 2006, 318p.

[Rap04] R. Rapp. “A freely available automatically generated thesaurus of related words”. In: Proceedings of the Forth Language Resources and Evaluation Conference – LREC’04, 2004, pp. 395-398.

[Rap08] R. Rapp. “The automatic generation of thesauri of related words for English, French, German, and Russian”. *International Journal of Speech Technology*, vol. 11-3&4, Dezembro 2008, pp. 147-156.

[RS68] P. M. Roget, B. Sears. “Thesaurus of English words: so classified and arranged as to facilitate the expression of ideas and assist in literary composition”. Gould and Lincol, 1868, 468p**.**

[Rug92] G. Ruge. “Experiments on linguistically-based term association”. *Information Processing & Management*, vol. 28-3, Janeiro 1992, pp. 317-332**.**

[Sah06] M. Sahlgren. “The Word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces”. Tese de doutorado, Department of Linguistics, Stockholm University, 2006, 156p.

[ST91] D.D.K. Sleator, D. Temperley. “Parsing English with a link grammar”. Technical report, Carnegie Mellon University Computer Science, 1991, 93p.

[ST08] A. Sumida, K. Torisawa. “Hacking wikipedia for hyponymy relation acquisition”. In: Proceedings of the Third International Joint Conference on Natural Language Processing - IJCNLP, 2008, pp. 883-888.

[SYY75] G. Salton, C.S. Yang, C.T. Yu. "A theory of term importance in automatic text analysis". *Journal of the American society for Information Science*, vol. 26-1, Jan-Fev 1975, pp. 33-44.

[Tan58] T.T. Tanimoto. “An elementary mathematical theory of classification”. Technical report, 1958, IBM Research, 238p.

[Une70] UNESCO. Guidelines for the establishment and development of monolingual thesauri. 1970, 68p.

[VSS+10] R. Vieira, D. da Silva, T. Sander, A. Augustini, C. Northfleet, F. Castilho, M. Bruckschen, P. Pizzinato, P. Bridi, P. Rao, R.L. Granada. “Representation and inference of privacy risks using semantic web technologies”. In: Proceedings of the EKAW2010 Poster and Demo Track - Volume 674, 2010.

[WCS96] L. Wall, T. Christiansen, R. Schwartz. "Programming Perl". Bonn: O'Reilly, 1996, 645p.

[WHZC09] P. Wang, J. Hu, H.J. Zeng, Z. Chen. “Using Wikipedia knowledge to improve text classification”. *Knowledge and Information Systems***,** vol. 19- 3, Maio 2009, pp. 265-281.

[WSG96] Y.A. Wilks, B.M. Slator, L.M. Guthrie. “Electric words: dictionaries, computers, and meanings”. MIT Press Cambridge, 1996, 301p.

[XY10] H. Xu, B. Yu. "Automatic thesaurus construction for spam filtering using revised back propagation neural network". *Expert Systems with Applications*, vol. 37-1, Janeiro 2010, pp. 18-23.

[YP05] D. Yang, D.M.W. Powers. “Measuring semantic similarity in the taxonomy of WordNet”. In: Proceedings of the Twenty-eighth Australasian Conference on Computer Science, 2005, pp. 315-322.

[YP08] D. Yang, D.M.W. Powers. “Automatic thesaurus construction”. In: Proceedings of the 31st Australasian conference on Computer science, 2008, pp. 147-156**.**

APÊNDICE A. AVALIADORES DO SISTEMA

Para a avaliação manual foram escolhidos três especialistas que conhecessem o domínio de privacidade, com perfis descritos abaixo.

- Avaliador 1:

O primeiro avaliador é um cientista pesquisador dos laboratórios da Hewlett Packard (HP)[8], em Princeton, Nova Jersey. É atualmente membro do Laboratório de Sistemas de Segurança da HP, que conduz pesquisas na área de privacidade de dados e segurança.

Antes de entrar para a HP, o mesmo trabalhou no laboratório STAR da empresa de tecnologias InterTrust[9], em Santa Clara, Califórnia em tópicos relacionados ao gerenciamento de direitos digitais (do inglês, *Digital Rights Management* - DRM). Atualmente seus interesses de pesquisa incluem privacidade, segurança computacional, criptografia e DRM. Ultimamente tem pesquisado e desenvolvido tecnologias que implementam as boas práticas de privacidade em grandes organizações.

- Avaliador 2:

O segundo avaliador também é um cientista pesquisador dos laboratórios da HP em Princeton, Nova Jersey. Atualmente é um estrategista de tecnologia da HP e tem seu foco principal em tecnologias de proteção e privacidade de dados. O avaliador é graduado em Ciências da Computação e em Direito, tendo mestrado em Direito na área de Leis de Governança Corporativas. Tem entre seus interesses principais a tecnologia, privacidade e leis.

- Avaliador 3:

O terceiro avaliador tem mais de treze anos de experiência na indústria de desenvolvimento de software e atualmente trabalha como pesquisador e projetista de software nos laboratórios da HP Brasil, em Porto Alegre. Desde fevereiro de 2008 trabalha com processos e iniciativas para o desenvolvimento de produtos que utilizam a segurança de dados. Atualmente o mesmo desempenha a função de líder em um programa de análise de informação.

[8] *http://www.hpl.hp.com/*

[9] *http://www.intertrust.com/star/*

APÊNDICE B. RESULTADO DA AVALIAÇÃO DOS TESAUROS

Este apêndice contém os resultados obtidos no processo de avaliação dos tesauros. Os resultados estão separados por avaliador, apresentando a posição do termo de acordo com a similaridade com o termo-chave (coluna *Pos.*), sendo esta classificação realizada pelo avaliador. Também é apresentado o tesauro de qual foi obtido o termo relacionado, o termo relacionado, e o julgamento do termo realizado. O julgamento (coluna *Av.*) é realizado como similar (indicado por *S*), não similar (indicado por *N*) e não tem certeza da similaridade do termo em relação ao seu termo-chave (indicado por *NS*).

Seed: *data_subject*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T5	*person*	1	*S*	1	*S*	2	*S*
T3,T5	*individual*	2	*S*	7	*S*	1	*S*
T3,T5	*user*	3	*S*	4	*S*	9	*S*
T4	*end_user*	4	*S*	3	*S*	8	*S*
T3,T5	*consumer*	5	*S*	2	*S*	5	*S*
T2	*customer*	6	*S*	8	*S*	6	*S*
T2,T3,T4,T5	*applicant*	7	*S*	17	*S*	16	*S*
T2,T3,T4,T5	*subscriber*	8	*S*	6	*S*	13	*S*
T3	*respondent*	9	*S*	-	*N*	12	*S*
T3	*complainant*	10	*S*	9	*S*	15	*S*
T1	*consent*	11	*S*	19	*S*	-	*N*
T1	*vital_interest*	12	*S*	-	*N*	-	*N*
T1	*processing*	-	*NS*	23	*S*	-	*N*
T4	*effective_authentication_method*	-	*NS*	-	*N*	-	*N*
T4	*calling_user*	-	*NS*	15	*S*	7	*S*
T4	*final_rule*	-	*NS*	-	*N*	-	*N*
T4	*calling_subscriber*	-	*NS*	14	*S*	11	*S*
T4	*applicable_state_authority*	-	*NS*	13	*S*	18	*S*
T5	*secretary*	-	*NS*	-	*N*	23	*S*
T2	*benefit*	-	*NS*	-	*N*	-	*N*
T3,T5	*employer*	-	*NS*	24	*S*	10	*S*
T2	*sending*	-	*N*	-	*N*	-	*N*
T2,T5	*commission*	-	*N*	-	*N*	3	*S*
T1	*data*	-	*N*	22	*S*	-	*N*
T4	*third_party_servicer*	-	*N*	16	*S*	24	*S*
T1	*erasure*	-	*N*	20	*S*	-	*N*
T1	*performance*	-	*N*	-	*N*	-	*N*
T1,T2,T3	*data_controller*	-	*N*	5	*S*	21	*S*
T1,T2	*controller*	-	*N*	12	*S*	22	*S*
T4	*following_term*	-	*N*	-	*N*	4	*S*
T2	*traffic_data*	-	*N*	18	*S*	-	*N*
T2	*financial_insititution*	-	*N*	11	*S*	17	*S*
T1	*third_party*	-	*N*	21	*S*	14	*S*
T5	*competent_authority*	-	*N*	25	*S*	20	*S*
T3	*regulatory_authority*	-	*N*	10	*S*	19	*S*
T1	*contract*	-	*N*	-	*N*	-	*N*

Seed: *children*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T1	*child*	1	*S*	2	*S*	1	*S*
T1	*age*	2	*S*	3	*S*	2	*S*
T2,T3	*school*	3	*S*	1	*S*	3	*S*
T3	*education*	4	*S*	5	*S*	6	*S*
T3	*citizen*	5	*S*	6	*S*	8	*S*
T1	*website*	6	*S*	-	*N*	-	*N*
T1	*commercial_website*	7	*S*	-	*N*	-	*N*
T4	*online_transaction*	8	*S*	-	*N*	-	*N*
T1	*target*	9	*S*	-	*N*	5	*S*
T1	*advertisement*	-	*NS*	-	*N*	10	*S*
T1	*violent*	-	*NS*	-	*N*	11	*S*
T5	*default*	-	*NS*	-	*N*	-	*N*
T1	*online_service*	-	*NS*	-	*N*	-	*N*
T1	*operator*	-	*NS*	-	*N*	-	*N*
T4	*setup_instruction*	-	*N*	-	*N*	-	*N*
T4	*marketing_script*	-	*N*	-	*N*	-	*N*
T4	*authority*	-	*N*	-	*N*	7	*S*
T3,T4	*directive*	-	*N*	-	*N*	9	*S*
T5	*drive*	-	*N*	-	*N*	-	*N*
T5	*vote*	-	*N*	-	*N*	-	*N*
T1	*practice*	-	*N*	-	*N*	-	*N*
T5	*generated*	-	*N*	-	*N*	-	*N*
T2	*official*	-	*N*	-	*N*	-	*N*
T2	*this_directive*	-	*N*	-	*N*	-	*N*
T4	*premium_payment*	-	*N*	-	*N*	-	*N*
T5	*anonymous*	-	*N*	-	*N*	-	*N*
T5	*height-power_limit*	-	*N*	-	*N*	-	*N*
T5	*religion*	-	*N*	-	*N*	-	*N*
T4	*promotional_material*	-	*N*	-	*N*	-	*N*
T2	*standard*	-	*N*	-	*N*	-	*N*
T2	*comments*	-	*N*	-	*N*	-	*N*
T2	*case*	-	*N*	-	*N*	-	*N*
T5	*development*	-	*N*	4	*S*	4	*S*
T5	*anticipation*	-	*N*	-	*N*	-	*N*
T5	*basic_power*	-	*N*	-	*N*	-	*N*
T4	*draft_investigation_report*	-	*N*	-	*N*	-	*N*
T3	*third_party*	-	*N*	-	*N*	-	*N*
T3	*protest*	-	*N*	-	*N*	12	*S*
T3	*consultation*	-	*N*	-	*N*	-	*N*
T3	*participation*	-	*N*	-	*N*	-	*N*
T3	*legal_person*	-	*N*	7	*S*	-	*N*
T3	*organization*	-	*N*	-	*N*	-	*N*
T4	*telephone_call*	-	*N*	-	*N*	-	*N*
T4	*comment*	-	*N*	-	*N*	-	*N*

Seed: *notice*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T3,T4,T5	*notification*	1	*S*	2	*S*	3	*S*
T1,T3,T4,T5	*statement*	2	*S*	14	*S*	4	*S*
T1	*initial_notice*	3	*S*	4	*S*	1	*S*
T1	*enforcement_notice*	4	*S*	19	*S*	2	*S*
T1	*writing*	5	*S*	16	*S*	-	*N*
T2	*principle*	6	*S*	-	*N*	-	*N*
T2	*regulation*	7	*S*	12	*S*	15	*S*
T1	*consumer*	8	*S*	15	*S*	-	*N*
T1	*opt*	9	*S*	-	*N*	9	*S*
T3	*copy*	10	*S*	-	*N*	-	*N*
T2	*law*	-	*NS*	13	*S*	16	*S*
T1,T3	*commissioner*	-	*NS*	18	*S*	-	*N*
T2	*term*	-	*NS*	10	*S*	11	*S*
T4	*confirmation*	-	*NS*	7	*N*	7	*S*
T2	*implementation*	-	*NS*	-	*N*	-	*N*
T2	*obligation*	-	*NS*	-	*N*	18	*S*
T1	*civil_infringement*	-	*NS*	-	*N*	17	*S*
T3,T5	*rule*	-	*NS*	-	*N*	12	*S*
T3,T5	*access*	-	*NS*	1	*S*	-	*N*
T5	*information*	-	*NS*	-	*N*	5	*S*
T5	*service*	-	*NS*	5	*S*	-	*N*
T3,T5	*code*	-	*NS*	-	*N*	-	*N*
T2	*sector*	-	*N*	-	*N*	-	*N*
T4	*checking*	-	*N*	-	*N*	-	*N*
T4	*avoidance*	-	*N*	-	*N*	-	*N*
T4	*option*	-	*N*	-	*N*	-	*N*
T4	*interrogatory*	-	*N*	-	*N*	-	*N*
T4	*officer*	-	*N*	-	*N*	-	*N*
T3,T5	*request*	-	*N*	6	*S*	-	*N*
T4	*mediation*	-	*N*	9	*S*	10	*S*
T5	*protection*	-	*N*	3	*S*	-	*N*
T2	*resource*	-	*N*	-	*N*	-	*N*
T4	*denial*	-	*N*	-	*N*	8	*S*
T5	*action*	-	*N*	-	*N*	6	*S*
T2	*interest*	-	*N*	11	*S*	14	*S*
T3	*form*	-	*N*	-	*N*	-	*N*
T1,T2	*subsection*	-	*N*	-	*N*	-	*N*
T3	*record*	-	*N*	8	*S*	13	*S*
T1	*appeal*	-	*N*	17	*S*	-	*N*

Seed: *marketing*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T2,T3,T4	*marketer*	1	*S*	1	*S*	3	*S*
T1	*joint_marketing*	2	*S*	5	*S*	1	*S*
T4	*contact*	3	*S*	7	*S*	-	*N*
T2	*address*	4	*S*	9	*S*	-	*N*
T3	*sender_information*	5	*S*	15	*S*	-	*N*
T5	*accurate_sender_information*	6	*S*	4	*S*	-	*N*
T1	*nonaffiliated*	7	*S*	14	*S*	-	*N*
T3	*consumer*	8	*S*	11	*S*	-	*N*
T1	*product*	9	*S*	12	*S*	7	*S*
T3	*sale*	10	*S*	10	*S*	8	*S*
T5	*purchase*	11	*S*	19	*S*	-	*N*
T1	*offering*	12	*S*	18	*S*	6	*S*
T1	*intermediary*	13	*S*	2	*S*	-	*N*
T1	*account_number*	14	*S*	-	*N*	-	*N*
T1	*third_party*	15	*S*	13	*S*	-	*N*
T1	*optout*	16	*S*	3	*S*	5	*S*
T1	*service*	-	*NS*	16	*S*	4	*S*
T4	*prohibit*	-	*NS*	-	*N*	-	*N*
T2	*cost*	-	*NS*	-	*N*	-	*N*
T4	*query*	-	*NS*	-	*N*	-	*N*
T1	*joint_agreement*	-	*NS*	17	*S*	-	*N*
T2	*code*	-	*N*	-	*N*	-	*N*
T2	*list*	-	*N*	8	*S*	-	*N*
T2	*development*	-	*N*	-	*N*	-	*N*
T2	*definition*	-	*N*	-	*N*	-	*N*
T2	*title*	-	*N*	-	*N*	-	*N*
T4	*competence*	-	*N*	-	*N*	-	*N*
T4	*satellite*	-	*N*	-	*N*	-	*N*
T4	*routing*	-	*N*	-	*N*	-	*N*
T4	*ownership*	-	*N*	-	*N*	-	*N*
T4	*connectivity*	-	*N*	-	*N*	-	*N*
T3,T4	*venture*	-	*N*	-	*N*	2	*S*
T5	*self-contained_device*	-	*N*	-	*N*	-	*N*
T5	*digital_form*	-	*N*	6	*S*	-	*N*
T5	*group_health_plan*	-	*N*	-	*N*	-	*N*
T5	*file_permission*	-	*N*	-	*N*	-	*N*
T5	*iris_configuration*	-	*N*	-	*N*	-	*N*
T3,T5	*following_information*	-	*N*	-	*N*	-	*N*
T5	*following_statement*	-	*N*	-	*N*	-	*N*
T5	*justification*	-	*N*	-	*N*	-	*N*
T3	*utilization*	-	*N*	-	*N*	-	*N*
T3	*file*	-	*N*	-	*N*	-	*N*
T1	*transaction_account*	-	*N*	-	*N*	-	*N*
T2	*loan*	-	*N*	-	*N*	-	*N*
T2	*delivery*	-	*N*	-	*N*	-	*N*
T1	*access_number*	-	*N*	-	*N*	-	*N*

Seed: *regulation*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T2,T3	*law*	1	*S*	15	*S*	2	*S*
T1	*statute*	2	*S*	1	*S*	8	*S*
T1	*order*	3	*S*	21	*S*	-	*N*
T1	*organic_law*	4	*S*	-	*N*	3	*S*
T3	*directive*	5	*S*	4	*S*	14	*S*
T3,T4	*clause*	6	*S*	3	*S*	-	*N*
T1	*provision*	7	*S*	13	*S*	-	*N*
T4	*principle*	8	*S*	5	*S*	12	*S*
T4	*subrule*	9	*S*	8	*S*	17	*S*
T3	*exemption*	10	*S*	16	*S*	22	*S*
T1	*interpretation*	11	*S*	17	*S*	15	*S*
T3,T4	*definition*	12	*S*	6	*S*	13	*S*
T5	*infringement_enforcement*	13	*S*	-	*N*	1	*S*
T2	*judge*	14	*S*	9	*S*	-	*N*
T3,T4	*restriction*	15	*S*	2	*S*	7	*S*
T3	*obligation*	16	*S*	-	*N*	5	*S*
T3	*condition*	17	*S*	-	*N*	16	*S*
T4	*brief*	18	*S*	-	*N*	-	*N*
T4	*sentence*	19	*S*	-	*N*	11	*S*
T1	*protection*	20	*S*	19	*S*	18	*S*
T2	*relief*	21	*S*	14	*S*	-	*N*
T2	*agreement*	22	*S*	10	*S*	4	*S*
T3	*standard*	23	*S*	11	*S*	10	*S*
T1	*consumer_credit_product*	-	*NS*	20	*S*	-	*N*
T1	*subchapter*	-	*NS*	22	*S*	-	*N*
T4	*checklist*	-	*NS*	-	*N*	19	*S*
T5	*new_paragraph*	-	*NS*	-	*N*	-	*N*
T5	*police_records_traffic_offence*	-	*NS*	-	*N*	-	*N*
T5	*site_operator*	-	*NS*	-	*N*	-	*N*
T3	*paragraph*	-	*NS*	-	*N*	9	*S*
T2	*title*	-	*NS*	12	*S*	-	*N*
T2	*location*	-	*N*	-	*N*	-	*N*
T1	*lending*	-	*N*	18	*S*	-	*N*
T1	*truth*	-	*N*	-	*N*	-	*N*
T4	*form*	-	*N*	-	*N*	-	*N*
T4	*scheme*	-	*N*	7	*S*	20	*S*
T5	*public_helpline_service*	-	*N*	-	*N*	-	*N*
T5	*box*	-	*N*	-	*N*	-	*N*
T5	*document_processing*	-	*N*	-	*N*	-	*N*
T5	*drive*	-	*N*	-	*N*	-	*N*
T5	*time_restriction*	-	*N*	-	*N*	6	*S*
T5	*interrelationship*	-	*N*	-	*N*	-	*N*
T2	*fee*	-	*N*	-	*N*	21	*S*
T2	*payment*	-	*N*	-	*N*	-	*N*

Seed: *customer*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T3,T4	*consumer*	1	*S*	2	*S*	2	*S*
T2,T3	*subscriber*	2	*S*	10	*S*	4	*S*
T2	*data_subject*	3	*S*	5	*S*	-	*N*
T3	*user*	4	*S*	11	*S*	9	*S*
T2,T3,T4	*individual*	5	*S*	7	*S*	7	*S*
T3,T5	*person*	6	*S*	1	*S*	6	*S*
T3	*recipient*	7	*S*	14	*S*	-	*N*
T3	*applicant*	8	*S*	9	*S*	11	*S*
T1	*notice*	9	*S*	18	*S*	-	*N*
T4	*licensee*	10	*S*	6	*S*	3	*S*
T2	*personal_data*	11	*S*	8	*S*	-	*N*
T1	*pii*	12	*S*	12	*S*	-	*N*
T1	*web_site*	-	*NS*	21	*S*	-	*N*
T1	*annual_notice*	-	*NS*	19	*S*	-	*N*
T4	*telephoned_party*	-	*NS*	4	*S*	10	*S*
T5	*joint_agreement*	-	*NS*	-	*N*	-	*N*
T5	*credit_reporter*	-	*NS*	-	*N*	5	*S*
T1	*loan*	-	*NS*	-	*N*	-	*N*
T1	*billing_statement*	-	*NS*	22	*S*	-	*N*
T2	*secretary*	-	*NS*	-	*N*	19	*S*
T5	*explicit_consent_experience*	-	*NS*	3	*S*	-	*N*
T1	*billing_entity*	-	*N*	20	*S*	1	*S*
T1	*government_authority*	-	*N*	-	*N*	15	*S*
T1	*governmental_entity*	-	*N*	-	*N*	16	*S*
T1	*donor*	-	*N*	-	*N*	-	*N*
T2	*damage*	-	*N*	-	*N*	-	*N*
T2	*representative*	-	*N*	15	*S*	8	*S*
T2	*controller*	-	*N*	16	*S*	17	*S*
T2	*license*	-	*N*	13	*S*	-	*N*
T4	*national_implementing_legislation*	-	*N*	-	*N*	-	*N*
T4	*wireless_network*	-	*N*	-	*N*	-	*N*
T4	*additional*	-	*N*	-	*N*	-	*N*
T4	*catalog*	-	*N*	-	*N*	-	*N*
T4	*compensating*	-	*N*	-	*N*	-	*N*
T4,T5	*telecommunications_service*	-	*N*	-	*N*	-	*N*
T5	*former_agency*	-	*N*	-	*N*	14	*S*
T5	*term_affiliation_period*	-	*N*	-	*N*	-	*N*
T2	*service*	-	*N*	17	*S*	-	*N*
T3,T5	*data_controller*	-	*N*	-	*N*	13	*S*
T3,T5	*regulatory_authority*	-	*N*	-	*N*	12	*S*
T5	*term_creditable_coverage*	-	*N*	-	*N*	-	*N*
T3	*employer*	-	*N*	-	*N*	18	*S*

Seed: *personal_data*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T3,T5	*pii*	1	*S*	2	*S*	1	*S*
T3,T5	*health_information*	2	*S*	4	*S*	4	*S*
T4	*identifier*	3	*S*	-	*N*	5	*S*
T3	*cardholder_data*	4	*S*	7	*S*	6	*S*
T4	*location_data*	5	*S*	6	*S*	8	*S*
T3,T4	*traffic_data*	6	*S*	-	*N*	11	*S*
T3,T5	*record*	7	*S*	1	*S*	7	*S*
T5	*person*	8	*S*	3	*S*	2	*S*
T4	*exempt_information*	9	*S*	10	*S*	-	*N*
T1	*data_subject*	10	*S*	24	*S*	13	*S*
T3,T5	*individual*	11	*S*	5	*S*	3	*S*
T1	*consumer*	12	*S*	19	*S*	-	*N*
T2,T3,T5	*information*	13	*S*	13	*S*	9	*S*
T1	*processing*	14	*S*	8	*S*	-	*N*
T1	*transfer*	15	*S*	22	*S*	-	*N*
T1,T5	*data_controller*	16	*S*	25	*S*	-	*N*
T1	*protection_of_individual*	17	*S*	26	*S*	-	*N*
T2	*customer_relationship*	18	*S*	-	*N*	-	*N*
T1,T3,T4,T5	*data*	19	*S*	18	*S*	14	*S*
T2	*opt*	-	*NS*	-	*N*	-	*N*
T2	*financial_product*	-	*NS*	15	*S*	-	*N*
T1	*cap*	-	*NS*	-	*N*	-	*N*
T1	*operator*	-	*NS*	23	*S*	-	*N*
T4	*electronic_media*	-	*NS*	11	*S*	-	*N*
T5	*regulatory_authority*	-	*NS*	9	*S*	-	*N*
T2	*reseller*	-	*NS*	17	*S*	-	*N*
T2	*technique*	-	*N*	-	*N*	-	*N*
T2	*term*	-	*N*	-	*N*	10	*S*
T2	*unit*	-	*N*	-	*N*	-	*N*
T1	*article*	-	*N*	21	*S*	-	*N*
T2	*audio_presentation*	-	*N*	-	*N*	-	*N*
T2	*complainant*	-	*N*	14	*S*	-	*N*
T4	*journalist*	-	*N*	-	*N*	-	*N*
T4	*manual_data*	-	*N*	-	*N*	12	*S*
T4	*draft*	-	*N*	-	*N*	-	*N*
T4,T5	*exclusive_application*	-	*N*	12	*S*	-	*N*
T3	*document*	-	*N*	20	*S*	15	*S*
T3	*list*	-	*N*	16	*S*	-	*N*

Seed: *personal_information*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T3,T5	*pii*	1	*S*	3	*S*	1	*S*
T4	*identifiable_information*	2	*S*	9	*S*	4	*S*
T3,T5	*nonpublic_personal_information*	3	*S*	1	*S*	3	*S*
T4	*sensitive_information*	4	*S*	8	*S*	2	*S*
T3,T5	*protected_health_information*	5	*S*	-	*N*	8	*S*
T3,T4,T5	*health_information*	6	*S*	7	*S*	9	*S*
T3	*credit_information*	7	*S*	4	*S*	11	*S*
T4	*patient_record*	8	*S*	11	*S*	7	*S*
T3,T5	*information*	9	*S*	2	*S*	18	*S*
T3,T5	*data*	10	*S*	5	*S*	19	*S*
T5	*individual*	11	*S*	6	*S*	6	*S*
T2	*collection*	12	*S*	15	*S*	-	*N*
T4	*requested_record*	13	*S*	-	*N*	-	*N*
T2	*access*	14	*S*	13	*S*	17	*S*
T2	*confidentiality*	15	*S*	10	*S*	16	*S*
T5	*vital_interest*	16	*S*	-	*N*	15	*S*
T1	*personal*	-	*NS*	19	*S*	5	*S*
T1	*authorized_recipient*	-	*NS*	18	*S*	10	*S*
T4	*users_desktop*	-	*NS*	-	*N*	14	*S*
T4	*such_information*	-	*NS*	-	*N*	-	*N*
T3	*record*	-	*NS*	12	*S*	20	*S*
T1	*ibm_web_site*	-	*N*	-	*N*	-	*N*
T1	*work_colleague*	-	*N*	-	*N*	12	*S*
T1	*terms_of_recommendation*	-	*N*	-	*N*	-	*N*
T1	*job_opening*	-	*N*	-	*N*	-	*N*
T1	*variety_of_situations*	-	*N*	-	*N*	-	*N*
T1	*member_country*	-	*N*	20	*S*	13	*S*
T1	*period*	-	*N*	-	*N*	-	*N*
T2	*statement*	-	*N*	-	*N*	-	*N*
T2	*obligation*	-	*N*	-	*N*	-	*N*
T2	*request*	-	*N*	17	*S*	-	*N*
T2	*description*	-	*N*	-	*N*	-	*N*
T2	*method*	-	*N*	-	*N*	-	*N*
T2	*respect*	-	*N*	14	*S*	-	*N*
T2	*communication*	-	*N*	-	*N*	-	*N*
T3,T4	*patient_safety_work_product*	-	*N*	-	*N*	-	*N*
T4	*disputed_amount*	-	*N*	-	*N*	-	*N*
T4	*post-sentence_condition*	-	*N*	-	*N*	-	*N*
T5	*youth*	-	*N*	-	*N*	-	*N*
T5	*financial_product*	-	*N*	-	*N*	-	*N*
T1	*ibm*	-	*N*	-	*N*	-	*N*
T3	*financial_institution*	-	*N*	16	*S*	-	*N*

Seed: *consent*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T1	*user*	1	*S*	-	*N*	-	*N*
T3,T4	*authorisation*	2	*S*	3	*S*	1	*S*
T3	*authorization*	3	*S*	4	*S*	2	*S*
T1	*processing*	4	*S*	18	*S*	-	*N*
T1	*data_subject*	5	*S*	20	*S*	-	*N*
T5	*person*	6	*S*	10	*S*	-	*N*
T2,T3	*agreement*	7	*S*	8	*S*	3	*S*
T1	*patient*	8	*S*	22	*S*	-	*N*
T3	*notification*	9	*S*	9	*S*	12	*S*
T1	*express*	10	*S*	17	*S*	-	*N*
T1	*subscriber*	11	*S*	24	*S*	13	*S*
T4	*checking*	-	*NS*	-	*N*	-	*N*
T5	*record*	-	*NS*	12	*S*	9	*S*
T5	*identification*	-	*NS*	-	*N*	-	*N*
T3,T5	*access*	-	*NS*	-	*N*	6	*S*
T3,T5	*right*	-	*NS*	1	*S*	5	*S*
T1	*organizer*	-	*N*	-	*N*	-	*N*
T1	*traffic_data*	-	*N*	19	*S*	-	*N*
T1	*accountholder*	-	*N*	16	*S*	-	*N*
T1,T5	*data*	-	*N*	15	*S*	-	*N*
T2	*table*	-	*N*	-	*N*	-	*N*
T2	*address*	-	*N*	-	*N*	-	*N*
T2	*privacy*	-	*N*	11	*S*	11	*S*
T2	*supervisor*	-	*N*	14	*S*	-	*N*
T2	*benefit*	-	*N*	-	*N*	8	*S*
T4	*coercion*	-	*N*	-	*N*	7	*S*
T4	*confirmation*	-	*N*	7	*S*	4	*S*
T4	*executing*	-	*N*	-	*N*	-	*N*
T2	*station*	-	*N*	-	*N*	-	*N*
T2	*organization*	-	*N*	-	*N*	-	*N*
T2	*controller*	-	*N*	13	*S*	-	*N*
T4	*desktop*	-	*N*	-	*N*	-	*N*
T4	*interrogatory*	-	*N*	-	*N*	-	*N*
T4	*encrypted*	-	*N*	-	*N*	-	*N*
T4	*eye*	-	*N*	-	*N*	-	*N*
T4	*revealment*	-	*N*	-	*N*	-	*N*
T3,T5	*request*	-	*N*	5	*S*	-	*N*
T5	*authority*	-	*N*	2	*S*	10	*S*
T5	*reasonable_opportunity*	-	*N*	-	*N*	-	*N*
T5	*information*	-	*N*	6	*S*	-	*N*
T3	*interest*	-	*N*	-	*N*	-	*N*
T2	*code*	-	*N*	-	*N*	-	*N*
T3	*identity*	-	*N*	21	*S*	-	*N*
T3	*representative*	-	*N*	23	*S*	-	*N*

Seed: *data_protection*							
		Avaliador 1		**Avaliador 2**		**Avaliador 3**	
Tesauro	**Termo**	**Pos.**	**Av.**	**Pos.**	**Av.**	**Pos.**	**Av.**
T1	*commission*	1	*S*	2	*S*	4	*S*
T1	*council*	2	*S*	-	*N*	5	*S*
T1	*principle*	3	*S*	-	*N*	1	*S*
T1	*commissioner*	4	*S*	5	*S*	2	*S*
T1	*european*	-	*NS*	1	*S*	-	*N*
T1	*schedule*	-	*N*	4	*S*	-	*N*
T1	*tribunal*	-	*N*	-	*N*	6	*S*
T1	*supervisor*	-	*N*	3	*S*	3	*S*
T1	*spanish*	-	*N*	-	*N*	-	*N*
T1	*entry*	-	*N*	-	*N*	-	*N*

ANEXO A. ETIQUETAS TREEBANK

Este anexo apresenta as etiquetas utilizadas pelo Penn Treebank e implementadas no analisador sintático desenvolvido em Stanford.

Etiqueta	Significado
CC	Coordinating conjunction - e.g. and, but, or..
CD	Cardinal Number
DT	Determiner
EX	Existential there
FW	Foreign Word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List Item Marker
MD	Modal e.g. can, could, might, may...
NN	Noun, singular or mass
NNP	Proper Noun, singular
NNPS	Proper Noun, plural
NNS	Noun, plural
PDT	Predeterminer - e.g. all, both ... when they precede an article
POS	Possessive Ending - e.g. Nouns ending in 's
PRP	Personal Pronoun - e.g. I, me, you, he...
PRP$	Possessive Pronoun - e.g. my, your, mine, yours...
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol - Should be used for mathematical, scientific or technical symbols
TO	to
UH	Interjection - e.g. uh, well, yes, my...
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner - e.g. which, and that when it is used as a relative pronoun
WP	Wh-pronoun - e.g. what, who, whom...
WP$	Possessive wh-pronoun
WRB	Wh-adverb - e.g. how, where why…