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

UM MODELO HÍBRIDO PARA O WSD EM BIOMEDICINA

RODRIGO RAFAEL VILLARREAL GOULART

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

Orientador: Vera Lúcia Strube de Lima

Porto Alegre
2013

Dados Internacionais de Catalogação na Publicação (CIP)

G694m Goulart, Rodrigo Rafael Villarreal
Um modelo híbrido para o WSD em biomedicina / Rodrigo Rafael Villarreal Goulart. Porto Alegre, 2013.
76 p.

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

1. Informática. 2. Semântica. 3. Linguística Computacional. 4. Algoritmos – Grafos. 5. Biomedicina. I. Lima, Vera Lúcia Strube de. II. Título.

CDD 006.35

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 TESE DE DOUTORADO

Tese intitulada "Um Modelo Híbrido para o WSD em Biomedicina", apresentada por Rodrigo Rafael Villarreal Goulart, como parte dos requisitos para obtenção do grau de Doutor em Ciência da Computação, Inteligência Computacional, aprovada em 26/03/2013 pela Comissão Examinadora:

Vera Lúcia Strube de Lima – PPGCC/PUCRS
Orientadora

Profa. Dra. Renata Vieira - PPGCC/PUCRS

Prof. Dr. Leandro Krug Wives - UFRGS

Prof. Dr. Thiago Alexandre Salgueiro Pardo – USP – São Carlos

Homologada em 11/06/2013, conforme Ata No. 010 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

DEDICATÓRIA

Dedico este trabalho aos meus pais, Dirceu e Norma, a minha esposa Melissa e aos meus filhos, Ana Clara e Pedro Henrique.

AGRADECIMENTOS

Muitas pessoas contribuíram para a realização desta tese de doutorado e não poderia deixar de manifestar meu profundo agradecimento a todas elas.
A PUCRS e DELL/Brasil pelo apoio financeiro para a realização deste trabalho.
Aos funcionários, alunos e professores do PPGCC pela convivência, amizade e experiências compartilhadas ao longo do curso de doutorado. Aos colegas e amigos do grupo de pesquisa pelo ótimo ambiente de trabalho e ótima convivência que me proporcionaram durante os anos que faço parte deste grupo.

UM MODELO HÍBRIDO PARA O WSD EM BIOMEDICINA

RESUMO

Este trabalho estuda o *Word Sense Disambiguation* no domínio da Biomedicina, para a língua inglesa, com uso de fontes externas de conhecimento. Dentre as propostas existentes para a seleção de um sentido para uma palavra ambígua, está a abordagem baseadas em grafos. Essa abordagem emprega uma métrica na avaliação de grafos que contêm candidatos ao sentido correto da palavra ambígua. Nesta pesquisa um conjunto de métricas é analisado individualmente e, com base nas avaliações, propõe-se um modelo híbrido de seleção de métricas com o objetivo de determinar a métrica mais adequada a ser empregada. O modelo faz uso de um conjunto de *features* e heurísticas que determinam uma solução semi-supervisionada para o WSD. Os resultados obtidos com experimentos apontam melhoria na performance e revelam novas perspectivas de pesquisa. O modelo proposto eleva a taxa de acerto a 68,48%, aumentando significativamente em 3,52% a taxa reportada na literatura.

Palavras Chave: *Word Sense Disambiguation*, Biomedicina, grafos, algoritmos, métricas.

A HYBRID MODEL FOR WSD IN BIOMEDICINE

ABSTRACT

This work studies Word Sense Disambiguation (WSD) in the Biomedicine domain for English language, using external knowledge sources. Among the existing proposals for the selection of a sense for an ambiguous word, there is the graph-based approach. This approach uses a metric in the evaluation of graphs containing candidates to the correct sense for the ambiguous word. In this research, a set of metrics is analyzed individually, and, based on this evaluation, we propose a hybrid model for the selection of the metrics in order to determine the most adequate metric to be employed. The model makes use of a set of features and heuristics that determine a semi-supervised solution for WSD. The results obtained with experiments show an improvement in performance and reveal new perspectives of research. The proposed model raises the hit rate to 68,48%, increasing significantly in 3,52% the rate reported in literature.

Keywords: Word Sense Disambiguation, Biomedicine, graphs, algorithms, metrics.

LISTA DE FIGURAS

LISTA TABELAS

SUMÁRIO

1. INTRODUÇÃO

A ambiguidade é o fenômeno linguístico em que uma palavra possui mais de um sentido. Ela representa um dos principais desafios para o Processamento da Linguagem Natural (PLN). O problema de *Word Sense Disambiguation* (WSD) na computação tem sido explorado desde 1950 e é considerado uma importante etapa no processamento de textos [1, 34]. Soluções para esse problema podem influenciar, por exemplo, a performance de sistemas para a tradução automática, mineração e classificação de textos.

Dentre as abordagens desenvolvidas ao longo desses anos estão aquelas com propósito específico, em que o WSD se restringe a um domínio do conhecimento. No domínio biomédico, sistemas que empregam PLN são projetados para analisar textos. A finalidade é indexar documentos e dar suporte à tomada de decisão. Para atingir esse objetivo esses sistemas devem lidar com ambiguidade. MedLEE[1] e PubMed[2] são dois exemplos. MedLEE é um sistema que extrai informações a partir de textos de radiologia. Ele organiza e classifica essas informações na forma de um vocabulário controlado. PubMed é um indexador de artigos biomédicos. Em ambos os casos, a busca por informação está associada à identificação e classificação dos conceitos presentes em textos.

No entanto, o processo de identificar automaticamente o sentido correto de uma palavra em um texto é um problema cuja solução ainda pode ser melhorada. Por exemplo, considere a busca da palavra *glucose* no indexador PubMed. De acordo com o metatesauro UMLS (Unified Medical Language System) [22], especializado na área de Biomedicina, a palavra *glucose* está presente em três conceitos: *glucose*, *plasma glucose measurement* e *glucose measurement*. O usuário que pesquisar pela palavra *glucose* no indexador PubMed pode desconhecer, e até mesmo não desejar, os resultados com os conceitos *plasma glucose measurement* e *glucose measurement*. Para identificar o sentido correto de uma palavra, o contexto em que ela foi empregada tem papel importante. Geralmente, conceitos ou simplesmente as palavras do entorno (i.e. palavras que ocorrem antes e depois da palavra ambígua em um texto) representam o contexto. Com esse tipo de informação é possível empregar algum método automático que considere a situação em que a palavra foi empregada, e então selecionar o sentido mais adequado de acordo com um conjunto pré-estabelecido de possíveis sentidos, como por exemplo aqueles estabelecidos no UMLS.

Abordagens baseadas no aprendizado supervisionado são comuns no WSD de textos de Biomedicina [24, 28, 47]. Contudo, elas exigem exemplos etiquetados para o treinamento, que podem estar ou não à disposição, ou serem de alto custo para serem elaborados. Essa limitação significa que as abordagens supervisionadas podem desambiguar uma amostra de palavras para a qual um conjunto de dados de treino foi elaborado, e isso limita sua utilização na prática. Por outro lado, abordagens não-supervisionadas não necessitam de exemplos etiquetados. Por fazerem uso de recursos estruturados como fonte de conhecimento, não há necessidade de um conjunto para treino e teste. Abordagens não-supervisionadas e semi-supervisionadas já foram exploradas anteriormente com o uso do UMLS [21, 29]. Além disso, essas fontes de conhecimento podem também ser utilizadas como um grafo, onde a topologia dessa estrutura de dados pode servir ao aprendizado não supervisionado. O UMLS, assim como a WordNet [31], estabelece relacionamentos semânticos entre os conceitos na forma de um grafo.

---

1 http://techventures.columbia.edu/news/columbia-grants-health-fidelity-exclusive-license-medlee-nlp (Último acesso: 28 de Fevereiro de 2013).

2 http://www.ncbi.nlm.nih.gov/pubmed/ (Último acesso: 28 de Fevereiro de 2013).

A estrutura de um grafo viabiliza o emprego de métricas para avaliação de vértices. Métricas no domínio da matemática são funções que generalizam a ideia geométrica de distância. No caso do UMLS, vértices correspondem a conceitos e as métricas estabelecem uma medida de importância para os mesmos. Existem algoritmos baseados em grafos para implementação dessas métricas. Entre os mais conhecidos, associados a recuperação de informações na Internet, estão o PageRank [8] e HITS [26]. Os algoritmos PageRank, Degree Centrality, Betweenness, Key Player Problem, entre outros, foram explorados em domínios especializados e não especializados [35, 2, 3]. Os resultados obtidos nessas pesquisas identificaram os melhores algoritmos para diferentes configurações de cenários (i.e. domínio do conhecimento, fonte de conhecimento, *corpora* para teste), mas lacunas ainda estão presentes.

Neste enquadramento este trabalho se propõe a um estudo dos problemas e soluções relacionados ao *Word Sense Disambiguation* (WSD) no domínio da Biomedicina. Este domínio é pesquisado pelo grupo de Processamento da Linguagem Natural da PUCRS, no qual esta Tese está inserida. Abordagens baseadas em grafos são exploradas com o objetivo de comparar, identificar lacunas e ampliar os resultados encontrados até o presente momento. Estudos preliminares demonstram que métodos não supervisionados com abordagens baseadas em grafos podem obter resultados semelhantes aos alcançados por métodos supervisionados. Este trabalho apresenta os resultados de experimentos não supervisionados e semisupervisionados baseados em grafos, que conduziram a um modelo híbrido para o processamento do WSD em Biomedicina.

No contexto do presente trabalho, métricas para grafos são métodos não supervisionados empregados no WSD. Estes métodos representam um **modelo simples** de processamento, por empregar uma métrica na seleção do sentido correto de uma ou mais palavras ambíguas. O estado da arte na pesquisa com o modelo simples apontam o PageRank Personalizado como a métrica mais indicada para o WSD no domínio da Biomedicina. A taxa de acerto dessa métrica é cerca de 66,16%. Contudo, os resultados de experimentos desenvolvidos nesta Tese revelam que existem casos em que outras métricas identificam o sentido de palavras corretamente, em casos que o PageRank Personalizado não o faz. Essa constatação revela que, se a métrica mais adequada para uma dada palavra ambígua for selecionada, as chances de sucesso na seleção, e na taxa de acerto, podem crescer significativamente.

A hipótese desta Tese é a de que é possível identificar a métrica mais adequada para uma dada palavra ambígua. Este trabalho apresenta uma proposta para implementar esse processo. Chamado de **modelo híbrido** de seleção de métricas para o WSD em Biomedicina para a língua inglesa, o objetivo desse modelo é estabelecer o processo de seleção da métrica mais adequada. Para isso são extraídas *features* da palavra ambígua e estas são empregadas, por heurísticas, na seleção da métrica. Experimentos foram elaborados para identificar quais *features* e heurísticas são as mais apropriadas para esse processo. O emprego de informações a respeito dos candidatos a sentido demonstra que este pode ser um meio (*feature*) adequado para identificar a métrica. A heurística que relaciona estas informações com a métrica é baseada na probabilidade condicional da ocorrência. Ela é estabelecida com base nos dados de experimentos com o modelo simples de WSD, reproduzidos neste trabalho a partir de pesquisas elaboradas por outros autores. O modelo híbrido constitui um modelo semisupervisionado para WSD. Os experimentos com o modelo proposto indicam uma taxa de acerto de 68,48%, melhorando significativamente, em 3,52%, os resultados do estado da arte relatado.

Este documento esta organizado da seguinte maneira. O Capítulo 2 resume os principais conceitos e métodos empregados no WSD. O Capítulo 3 apresenta as principais abordagens baseadas em grafos para o WSD. São descritos os meios para representar grafos, algoritmos e resultados obtidos por outros autores. O emprego da abordagem baseada em grafos no domínio da Biomedicina é o tema do Capítulo 4. *Corpora* e propostas de WSD supervisionadas e não-

supervisionadas não descritos, assim como os resultados obtidos até então com essas abordagens. O Capítulo 5 apresenta o emprego individual das métricas, e os resultados obtidos individualmente por três métricas. Lacunas são identificadas e discutidas. O Capítulo 6 apresenta o modelo híbrido de seleção de métricas. As alternativas de *features* e heurísticas para seleção de métricas são analisadas também neste capítulo. Por fim, o Capítulo 7 apresenta conclusões e sugestões de trabalhos futuros.

2. *WORD SENSE DISAMBIGUATION*

O processo de selecionar o sentido correto de uma palavra é chamado ***Word Sense Disambiguation***. Identificar sentidos de palavras auxilia o aperfeiçoamento de outras áreas de aplicação do Processamento da Linguagem Natural. Tradução automática, sistemas para perguntas e respostas, recuperação de informações e a classificação de textos também são exemplos de aplicações desse processo. A maneira como o WSD é explorado nessas e outras áreas de aplicação varia de acordo com as suas particularidades. A discussão apresentada aqui ignora essas diferenças específicas e foca no WSD como uma área independente.

Na forma mais elementar, algoritmos para o WSD consideram como entrada palavras com uma lista fixa de potenciais sentidos. Como resultado, retorna a palavra que representa o sentido correto para um determinado emprego. A natureza da entrada e do inventário de sentidos depende da área de aplicação. Para tradução automática do Inglês para o Português, o inventário de etiquetas de sentidos para uma palavra em Inglês será um conjunto de diferentes traduções da mesma para o Português. Se o objetivo é a sintetização de voz, o inventário deve se restringir a homógrafos com diferentes pronúncias como no caso das palavras "colher", "seca" e "jogo" como substantivos ou verbos. Se o objetivo é a indexação de artigos biomédicos, um exemplo é o inventário de sentidos e etiquetas, como o tesauro MeSH (*Medical Subject Headings*)[3]. Quando tratado isoladamente o WSD, é possível utilizar dicionários ou tesauros como WordNet ou LDOCE [25].

Existem duas variantes do processo genérico de WSD [25, 34]. A primeira é o processamento ***lexical sample***, onde um pequeno conjunto de palavras pré-selecionadas é escolhido, juntamente com um inventário de sentidos para cada palavra encontrada em algum dicionário. Como o conjunto de palavras e o conjunto de sentidos são pequenos, abordagens baseadas em aprendizado de máquina são geralmente empregadas no processamento *lexical sample*. Para cada palavra, um número de instâncias do *corpus* (sentenças do contexto) podem ser selecionadas e etiquetadas manualmente com o sentido correto de cada uma das palavras em análise. Sistemas de classificação podem então ser treinados com esses exemplos etiquetados. Palavras não etiquetadas podem ser então etiquetadas com a utilização do classificador treinado. Trabalhos anteriores em WSD empregavam exclusivamente esse método no processamento *lexical sample*, construindo algoritmos específicos para a desambiguação de uma única palavra.

Por outro lado, no processamento ***all-words****,* um programa recebe textos inteiros e um dicionário com um inventário de sentidos para cada entrada, com a tarefa de desambiguar cada palavra contida no texto. O processamento *all-words* é similar ao da anotação morfossintática, exceto por considerar um conjunto muito maior de etiquetas já que cada palavra tem seu próprio conjunto de sentidos. A consequência desse grande conjunto de etiquetas é um sério problema de dados esparsos, pois é improvável produzir dados de treino para cada palavra presente no conjunto de teste. Além disso, considerando o possível número de palavras polissêmicas presentes em um dicionário comum, as abordagens baseadas no treinamento de um classificador por termo são impraticáveis.

3 http://www.nlm.nih.gov/mesh/. Último acesso 4 de Agosto de 2012.

2.1. Seleção de Sentidos

Um **sentido** é um significado comum para uma palavra. Por exemplo, considere as seguintes sentenças:

(a) A manga é indicada para o tratamento de anemia.
(b) Manga Doce é o melhor restaurante do país.
(c) Entre 1969 e 1974, Manga foi muitas vezes campeão nacional.

A palavra "manga" é utilizada nas sentenças com três sentidos diferentes: a fruta (a), um restaurante (b) e o jogador de futebol (c). Um **inventário de sentidos** tem como objetivo relacionar um intervalo finito de significados de uma palavra. Contudo, a elaboração de um inventário não é uma tarefa trivial. Os exemplos (b) e (c) podem se referir a dois campeões, mas cada um deles se refere a uma categoria diferente (gastronomia e esporte, respectivamente). Por outro lado, o nível de especificidade que se utiliza pode depender do tipo de aplicação. Por exemplo, para um sistema de recuperação de informações, pode ser relevante associar "manga" ao conceito de um "jogador" e, mais especificamente, um "goleiro".

Existem duas abordagens comuns para tornar explícitos os sentidos de uma palavra. A primeira é a abordagem **enumerativa**, que relaciona sentenças descrevendo o sentido. A Figura 2.2 exemplifica entradas para o substantivo "manga" presente nas sentenças anteriores. Contudo, esta abordagem é limitada, se for necessário organizar os sentidos em termos de especificidade e generalidade.

manga s. **1.** fruta da mangueira **2.** restaurante temático da cidade do Rio de Janeiro **3.** famoso goleiro de futebol brasileiro.

Figura 2.1: Exemplo de enumeração de sentidos

A abordagem gerativa apresentada por Pustejosky [45] se propõe, por outro lado, a estabelecer os sentidos de uma palavra através do seu relacionamento com outras palavras a partir de regras. A instanciação de um conjunto de regras leva à criação de sentidos de uma dada palavra. A Figura 2.1 apresenta um exemplo simples para a definição da palavra *bank* [45]. Nele estão relacionadas duas variantes de sentido da palavra, diferenciadas por dois aspectos: categoria lexical (CAT) e classe (GENUS), que associa a palavra com alguma taxonomia pré-estabelecida.

$$\begin{bmatrix} \textbf{bank} \\ \text{CAT} = \textbf{count\_noun} \\ \text{GENUS} = \textbf{financial\_institution} \end{bmatrix}$$

$$\begin{bmatrix} \textbf{bank} \\ \text{CAT} = \textbf{count\_noun} \\ \text{GENUS} = \textbf{shore} \end{bmatrix}$$

Figura 2.2: Exemplo de uma definição gerativa para a palavra *bank* [45]

De acordo com Navigli [34], a abordagem enumerativa é a mais adotada pela comunidade científica, e por essa razão será adotada neste trabalho.

2.2. Fontes Externas de Conhecimento

Fontes de conhecimento fornecem dados que são essenciais para associar sentidos às palavras. Por essa razão, conhecimento é fundamental para o WSD. *Corpora* anotado, ontologias e tesauros são exemplos de fontes de conhecimento e estas podem ser classificadas como recursos estruturados ou recursos não estruturados.

Recursos estruturados são aqueles que estabelecem conjuntos de conceitos e relacionamentos semânticos com a finalidade de viabilizar o seu processamento automatizado. Tesauros e ontologias são exemplos desse tipo de recurso [34].

Tesauros fornecem informações sobre o relacionamento entre palavras, como sinonímia (e.g. "moto" é sinônimo de "motocicleta") e antonímia (e.g. "bom" é antônimo de "mau"), entre outras relações, como uma medida de similaridade ou distância semântica [25]. Duas palavras são mais similares se elas compartilham mais significados ou são quase sinônimas. De forma inversa, são menos similares ou mais distantes semanticamente as palavras que possuem menos sentidos em comum. Desta forma, os relacionamentos de sinonímia e similaridade definem relações entre sentidos ao invés de relações entre palavras.

Outro exemplo de recursos estruturados são as **ontologias**. Estas são a "especificação explícita de uma conceitualização" [18], que se refere a um conjunto de distintos conceitos de um único domínio [25]. WordNet [31] e UMLS [22] são considerados ontologias por estabelecerem uma rede de relacionamentos semânticos entre conceitos de domínio geral (WordNet) e específico (UMLS).

A WordNet é uma base dados lexical destinada ao uso computacional. Substantivos, verbos, adjetivos e advérbios estão organizados na forma de conjuntos de sinônimos, chamados *synsets*, cada um representando um conceito. Além da relação de sinonímia, outras relações semânticas estão presentes na WordNet, como hiponímia e hiperonímia. A WordNet, na sua mais recente versão (3.1), contém cerca de 155.000 palavras relacionadas a cerca de 117.000 *synsets*. Por exemplo, a palavra *adjustment* é expressa na WordNet com o seguinte *synset*:

$$\{adjustment_s^1, accommodation_s^1, fitting_{adj}^1\}$$

Este *synset* representa um conjunto de palavras com sentido semelhante. Nele, o valor subscrito de cada conceito representa uma etiqueta morfossintática para substantivos, verbos, adjetivos e advérbios (aqui indicada como s, v, adj e adv, respectivamente). Os valores sobrescritos representam o *sense number,* que é uma espécie de identificador único do *synset*. Assim, a palavra *adjustment* possui diversos significados e estes representam um conjunto de *synsets* definidos como:

$$\{\{adjustment_s^1, accommodation_s^1, fitting_{adj}^1\},$$
$$\{adjustment_s^2, alteration_s^2, modification_s^1\},$$
$$\{adjustment_s^3, registration_s^5, readjustment_s^2\},$$
$$\{adjustment_s^4, adaptation_s^1, adaptation_s^2\},$$
$$\{adjustment_s^5, allowance_s^3\}\}$$

O UMLS, descrito no Capítulo 4, é considerado um metatesauro por representar a unificação de um amplo conjunto de vocabulários controlados de medicina, além de sistemas de classificação. O metatesauro é organizado com base em conceitos e cada um é relacionado a um *Concept Unique Identifier* (CUI). Por exemplo, os seguintes CUIs estão associados ao termo *cold*:

{{C0009443,'Common Cold'},
{C0009264 'Cold Temperature'},
{C0234192 'Cold Sensation'}}

O metatesauro também contém informações sobre as relações entre CUIs, expressas na forma de bases de dados em tabelas. A tabela MRREL reúne as relações entre CUIs em diferentes tipos. Por exemplo, C0009443 *Common Cold* está relacionado com o C0035243 *Respiratory Tract Infections* pela relação chamada PAR (*parent*). Outros tipos de relações na tabela MRREL incluem QB (pode ser qualificado por - *qualified by*), RQ (relacionado e possivelmente sinônimo) e RO (relacionado, outro). Por exemplo, {C0009443,'*Common Cold*'} está relacionado com {C0460004 '*Head and Neck*'} pela relação RO. A tabela MRREL também contém a fonte onde a relação foi obtida. Por exemplo, C0009443 e C0460004 são encontrados no *National Cancer Institute Thesaurus*. A mesma relação pode ser encontrada em múltiplas fontes, por exemplo as CUIs C0009443 e C0035243 são encontradas em quatro diferentes fontes.

As relações de coocorrência entre CUIs são encontradas na tabela MRCOC. Essas relações ocorrem entre conceitos similares (e.g. {C00004238,'*Artial Fibrillation*'} e {C0003811 '*Cardiac Arrhythmia*'}) ou diferentes conceitos que compartilham relações importantes (e.g. {C00004238,'*Artial Fibrillation*'} e {C0012265,'*Digoxin*'}). Apesar de a tabela MRCOC incluir um grande número de relações de coocorrência, a maioria dos conceitos não tem qualquer relação de coocorrência associada. A associação nessa tabela foi criada automaticamente pelo processamento de três fontes de informação: MEDLINE, 2002-2007; AI/RHEUM, 1993 e o *Canonical Clinical Problem Statement System*, 1999. A tabela MRCOC inclui detalhes sobre o peso da relação de coocorrência entre conceitos baseado no número de coocorrências identificadas [3].

Recursos não estruturados, por sua vez, são aqueles que reúnem palavras sem estabelecer qualquer repositório explícito exclusivo de conceitos e seus relacionamentos semânticos. *Corpora*, anotados ou não anotados, são exemplos desse tipo de recurso. Além de *corpora*, informações contidas indiretamente nos textos também são utilizadas por métodos de classificação em abordagens supervisionadas e não-supervisionadas (Seção 2.4). Listas com a coocorrência de palavras, listas de frequência de palavras e *stoplists* (tais como listas de palavras de uso comum) também são utilizadas como recursos não estruturados para o WSD.

2.3. Representação do Contexto

Texto é considerado uma fonte de informação não estruturada. Para poder processá-lo algoritmicamente, é necessário convertê-lo numa forma estruturada, o que geralmente é feito por meio do **pré-processamento**. Essa etapa costuma incluir as seguintes subetapas: tokenização, anotação morfossintática, lematização, *chunking* e *parsing*.

A **tokenização** é uma etapa de normalização que divide o texto em *tokens*, geralmente associados a palavras. A **anotação morfossintática** consiste na atribuição de uma categoria gramatical para cada palavra, simples ou composta. Considere por exemplo a seguinte frase anotada: "A/DET manga/S é/V indicada/V para/PRP o/DET tratamento/S de/PRP anemia/S.". As

etiquetas DET, S, V e PRP representam artigos definidos, substantivos, verbos e preposições, respectivamente. A *lematização* é a redução das variantes morfológicas de uma palavra a uma forma base, seu lema. Por exemplo, "é → ser", "indicada → indicar" e "a → o". **Chunking** é processo de divisão de um texto em partes sintaticamente relacionadas. Por exemplo, [A manga]$_{SN}$ [é indicada]$_{SV}$ [para o tratamento de anemia]$_{SP}$, onde SN, SV e SP significam sintagma nominal, sintagma verbal e sintagma preposicional, respectivamente. Por fim, o **parsing** é o processo de identificação da estrutura de uma sentença, com a geração de uma árvore que a representa. Como resultado do pré-processamento da mesma sentença, temos a sequência representada na Figura 2.3.

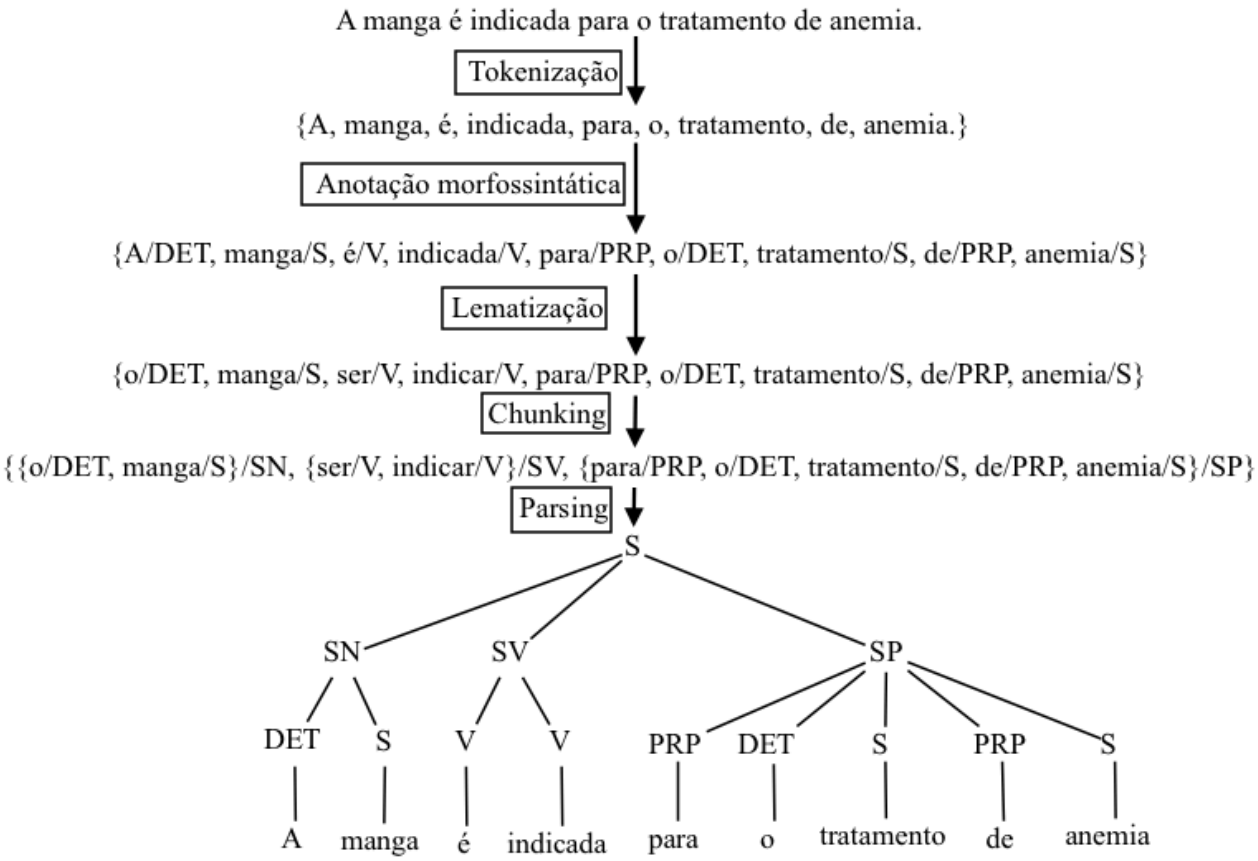

Figura 2.3: Etapas no pré-processamento de textos

O resultado de cada uma dessas etapas pode ser retratado como um vetor de caracteres. Computacionalmente, essas informações podem ser então utilizadas e complementadas com mais informações, por exemplo informações a respeito do contexto. Assim, métodos automáticos podem encontrar o sentido mais apropriado para uma palavra (e.g. "manga"), presente num inventário de sentidos, por intermédio destas e outras informações a respeito do contexto da palavra ambígua.

Estas informações a respeito do contexto são chamadas de *features* ou atributos e incluem as informações identificadas em cada um dos passos do pré-processamento. Mas podem incluir outras, como a frequência de uma palavra ou relações de sinonímia, por exemplo. Segundo Navigli [34] as *features* podem ser reunidas nos seguintes grupos:

- ***features* locais** representam o contexto local em que uma palavra foi empregada, tal como a categoria morfossintática ou lema de um pequeno intervalo de palavras no entorno da palavra em análise;

- ***features* de tópico** definem, ao contrário das *features* locais, o tópico geral de um texto ou porção do discurso (referente a um conjunto de palavras, como uma sentença ou frase, por exemplo);
- ***features* sintáticas** indicam relações sintáticas entre a palavra ambígua e outras presentes na mesma sentença;
- ***features* semânticas** expressam informações semânticas, como o sentido das palavras dentro do contexto, indicadores do domínio das palavras, etc.

Com as *features* estabelecidas, cada ocorrência de uma palavra (geralmente dentro de uma sentença) pode ser convertida num vetor. Por exemplo, a Figura 2.4 ilustra exemplos simples de vetores de *features* para as sentenças (a) e (b) apresentadas na Seção 2.1. Nele estão indicadas as categorias morfossintáticas de quatro palavras do contexto da palavra "manga", as duas anteriores e as duas posteriores. Também está relacionado o sentido correto da palavra "manga" nesse contexto.

Sentença	w_{-2}	w_{-1}	w_{+1}	w_{+2}	Sentido
(a)	-	Artigo definido	Verbo	Verbo	FRUTA
(b)	-	-	Adjetivo	Verbo	RESTAURANTE

Figura 2.4: Exemplos de vetores de features

A extensão do contexto pode ser estabelecida de diferentes formas. Unigramas, bigramas, trigramas ou mesmo uma janela de palavras, como é proposto na Figura 2.4, são formas de estabelecer o intervalo de palavras do contexto. Utilizar palavras de uma determinada categoria morfossintática é outra forma de estabelecer o vetor. Por exemplo, podemos selecionar todos os substantivos presentes em uma janela de palavras.

Com as informações do contexto representadas de forma estruturada, como a de um vetor, é possível aplicar métodos de classificação na tentativa de identificar o sentido das palavras ambíguas.

2.4. Método de Classificação

A maioria das abordagens de resolução da ambiguidade de palavras tem origem no aprendizado de máquina, variando entre métodos fortemente supervisionados e abordagens de reconhecimento de padrões estruturais e sintáticos.

A **classificação supervisionada** no WSD emprega técnicas de aprendizado de máquina na construção de um classificador a partir de um conjunto de treino anotado. Os exemplos são codificados com um conjunto de *features* e o sentido correto da palavra ambígua. Por outro lado, a **classificação não-supervisionada** no WSD utiliza *corpora* não etiquetados ou quaisquer outros *corpora* anotados com sentidos. Além disso, é necessário distinguir as abordagens baseadas em conhecimento das abordagens baseadas em *corpus*, ou mais pobres em conhecimento. As abordagens **baseadas em conhecimento** utilizam recursos externos, como dicionários, tesauros e ontologias. Esses recursos não representam informações linguísticas, mas informações a respeito do domínio das palavras que se deseja analisar (por exemplo, tesauro UMLS empregado na desambiguação de sentidos de termos em textos de Biomedicina). As abordagens **baseadas em *corpus***, por outro lado, não utilizam recursos externos para desambiguação.

A combinação de métodos supervisionados ou não, com abordagens baseadas em conhecimento ou *corpus*, segundo Navigli [34], pode variar e trazer diferentes resultados. A maior parte das abordagens baseadas em conhecimento que emprega propriedades estruturais, como a estrutura de grafos em redes semânticas, utiliza mais supervisão e conhecimento do que aqueles baseados apenas em sobreposição de grafos ou métodos para determinar a dominância de um sentido.

Por fim, abordagens para o WSD podem ser classificadas como baseadas em *tokens* (*token-based*) ou baseadas em um padrão, chamadas *type-based*. A abordagem ***token-based*** associa um sentido específico para cada ocorrência de uma palavra ambígua em um texto. Isso significa que, dependendo do contexto em que se encontra, uma palavra pode assumir diferentes significados. Por outro lado, a desambiguação ***type-based*** se baseia na hipótese de que uma palavra é utilizada com o mesmo sentido ao longo de um mesmo texto. Consequentemente, estas abordagem podem inferir o sentido predominante de uma palavra, para então utilizá-lo em todas as ocorrências da palavra.

3. ABORDAGENS DE GRAFOS PARA O WSD

Nos últimos anos há um crescente interesse pela pesquisa sobre redes, sejam elas sociais, de hiperdocumentos ou outras. Em parte devido ao crescimento da Internet, mas também devido ao desenvolvimento de algoritmos para análise de *links* com objetivo de recuperar informações. Entre eles, PageRank [8] e HITS [26] são os mais lembrados. No WSD, os recursos estruturados podem ser utilizados com esse tipo de algoritmo como um meio de classificar os termos candidatos a desambiguação.

O PageRank determina um peso para cada elemento de um conjunto de documentos correlacionados por meio de *hyperlinks*. O objetivo é medir a importância relativa de um documento dentro do conjunto de documentos. HITS classifica páginas pelos seus valores de autoridade e centralidade. Enquanto um documento central aponta para outros documentos com distâncias mínimas, um documento com autoridade é aquele para o qual muitos apontam. Além da recuperação de informação, algoritmos para análise de *links* foram empregados em outras tarefas, como a detecção de *spam* [19], recuperação de informação por tópicos [11], busca por palavra-chave em bases de dados relacionais [5], e medida de fator de impacto [6].

Abordagens baseadas em grafos têm se popularizado na área do PLN. A razão para tal, em muitos casos, é que vários problemas estão associados à seleção do melhor candidato à solução, dentre vários outros inter-relacionados. A desambiguação do sentido de um palavra é um exemplo de problema. Considerando a viabilidade de um dicionário relacionar palavras, seus diferentes sentidos e relacionamentos com as demais palavras em um domínio, é possível estabelecer métodos para tentar determinar o sentido correto. Ou seja, quando uma palavra ambígua é empregada em uma sentença, o resultado da busca nesse dicionário é composto por múltiplas entradas. A análise dos relacionamentos de cada uma delas com as demais pode subsidiar a escolha de uma das entradas para então determinar o sentido correto. As diferentes interpretações de uma palavra podem ser representadas de forma compacta como um grafo, onde os nodos correspondem a sentidos e as arestas a relacionamentos entre os sentidos (e.g. sinonímia, hiponímia etc.). Desta forma, a tarefa é determinar um único sentido para uma palavra ambígua dentro do contexto. Isso pode ser feito, por exemplo, com a seleção do sentido com maior número de conexões (i.e. arestas de entrada) no grafo [17]. Estes relacionamentos podem receber pesos de acordo com o seu tipo semântico. Por exemplo, relações de sinonímia podem ser mais importantes que as de hiponímia. Navigli e Velardi [38] classificam os sentidos de uma palavra de acordo com a distância entre os nodos do grafo. O PageRank também é empregado na classificação dos sentidos de uma palavra ambígua [3]. Algoritmos para grafos são considerados ideais para o WSD por serem não supervisionados e, por essa razão, não necessitarem de dados anotados manualmente com os sentidos corretos.

Além do WSD, algoritmos para grafos foram empregados nas tarefas de sumarização [52], análise de sentimentos [13], recuperação de sentenças em sistemas de pergunta e resposta [41], aprendizagem de ontologias [37], e análise de diálogos [14].

Apesar da existência de métodos baseados em grafos para o PLN, existem poucos estudos que apresentem um levantamento de como o grau de conectividade de grafos, e as diferentes formas de mensurá-lo, podem afetar diferentes tarefas. Métricas para conectividade de grafos têm sido propostas na análise de redes sociais e aplicadas a diferentes tipos de redes. Newman [39] discute a dificuldade em estabelecer uma métrica universal. Para Newman, identificar quais são as propriedades mais importantes de um grafo é uma tarefa fortemente atrelada às respostas que se deseja extrair do grafo. Trabalhos com a abordagem baseada em grafos utilizaram quase

exclusivamente duas alternativas e suas variantes: grau de centralidade e PageRank. As métricas baseadas em similaridade (entre a palavra ambígua e as outras do contexto), por outro lado, foram avaliadas por outros trabalhos em WSD [29, 21]. Outra questão importante é o dicionário empregado na construção do grafo de sentidos. Ele determina a topologia do grafo e determina seus padrões de conexões. Por exemplo, um grafo densamente conectado será criado a partir de um dicionário com muitas relações de sentido. O trabalho de Navigli e Lapata [36] explora essas questões. Esse trabalho emprega duas versões da WordNet em experimentos com nove métricas de conectividade para o WSD. Contudo, os resultados não deixam claro se as diferenças de performance no WSD estão relacionadas a uma métrica de conectividade, a um processo de construção de um grafo, ou se recebem influência de ambos [35]. A Seção 3.3 apresenta estes e outros resultados.

3.1. Utilizando Grafos a Partir de Textos

A utilização de algoritmos baseados em grafos passa por duas etapas. A primeira é a construção de um subgrafo dos conceitos presentes em uma base de conhecimento lexical pré-estabelecida. Ele relaciona um candidato a sentido de uma palavra ambígua com as palavras encontradas no contexto. Por exemplo, considere o seguinte parágrafo:

> *... and the regression coefficient of percentage decline in FEV1 with log dose, were calculated ("slope", after transformation), with and without calibration of nebulizers by weight and* ***adjustment*** *for nonresponse bias. Standardization for baseline lung function and variation in smoking prevalence was applied to slope. Results were ...*

Este parágrafo foi extraído do *corpus* NLM-WSD proposto em [53], apresentado na Seção 4.1. O *corpus* inclui 5000 textos contendo 50 palavras ambíguas anotadas. Cada anotação tem 100 instâncias (textos). As textos são resumos extraídos aleatoriamente da base MEDLINE em 1998. As instâncias foram manualmente desambiguadas por 11 anotadores, que anotaram cada ocorrência do termo com o significado correspondente encontrado no UMLS. A palavra *adjustment*, por exemplo, tem três sentidos possíveis, indicados no *corpus* como:

1. Individual Adjustment
2. Adjustment Action
3. Psychological adjustment

Neste caso, a primeira opção representa o sentido escolhido pelos anotadores. Algumas instâncias foram classificadas como *none* para indicar que os anotadores não encontraram um possível significado para o termo no UMLS. Cerca de 1017 instâncias, ou 20,34%, foram classificadas como *none*, o que já denota as dificuldades, mesmo humanas, nessa área.

Para construir o grafo que represente os termos presentes no contexto da palavra ambígua, os demais conceitos presentes no contexto devem ser identificados. Considerando uma janela de 20 termos, 10 antes da palavra ambígua e 10 após, temos a seguinte anotação:

> *and the regression coefficient of [percentage]$_{-10}$ decline in [FEV1] $_{-9}$ with [log]$_{-8}$ [dose]$_{-7}$, were [calculated]$_{-6}$ ("[slope]$_{-5}$", after [transformation]$_{-4}$), with and without [calibration]$_{-3}$ of [nebulizers]$_{-2}$ by [weight]$_{-1}$ and* ***[adjustment]$_0$*** *for nonresponse [bias]$_{+1}$. [Standardization] $_{+2}$ for [baseline] $_{+3}$ [lung function]$_{+4}$ and [variation]$_{+5}$ in [smokin]$_{+6}$ [prevalence]$_{+7}$ was [applied]$_{+8}$ to [slope] $_{+9}$. [Results]$_{+10}$ were ...*

As palavras entre colchetes determinam os conceitos e a sua posição relativa ao termo ambíguo. Por exemplo, a sexta palavra antes do termo ambíguo é *calculated*. Termos compostos podem ser encontrados (ex.: [*lung function*]$_{+4}$). Como os termos do contexto são aqueles encontrados no metatesauro UMLS, os mesmos podem ser ambíguos, como é o caso de [*variation*]$_{+5}$. Nesta situação, o primeiro sentido encontrado é o utilizado neste exemplo.

De uma forma geral os algoritmos para desambiguação de sentidos procedem de forma incremental, uma sentença σ por vez. Assim, inicialmente é construído um grafo $G = (V, A)$ para cada sentença, que é induzido a partir do grafo do léxico de referência (base de conhecimento). Os vértices do grafo são sentidos de palavras e as arestas são relações semânticas. Para cada palavra $w_i \in \sigma$, temos o conjunto de sentidos de w_i presente no léxico utilizado, definido como $Sentidos(w_i)$, e o sentido mais apropriado para w_i definido como $S_{wi} \in Sentidos(w_i)$.

A construção do grafo G a partir do léxico de referência segue um conjunto de passos, como definidos em [36]. Considerando uma sequência de palavras $\sigma = (w_1, w_2, ..., w_n)$, temos:

1. Considere $V_\sigma \coloneqq \bigcup_{i=1}^{n} Sentidos(w_i)$ determina o conjunto de todos os sentidos possíveis em σ. Determinamos então que $V \coloneqq V_\sigma$ e $A \coloneqq \varnothing$.
2. Para cada vértice $v \in V_\sigma$ uma busca em profundidade (*Depth-First-Search*) é feita no grafo do léxico de referência. Quando um vértice $v' \in V_\sigma (v' \neq v)$ é encontrado por intermédio de um caminho $v, v_1, v_2, ..., v_k, v'$ de tamanho L, todos os vértices e arestas intermediários desse caminho são adicionados ao grafo G. Assim, $V \coloneqq V \cup \{v_1, ..., v_k\}$ e $A \coloneqq A \cup \{\{v, v_1\}, ..., \{v_k, v'\}\}$.

Para exemplificar o processo de construção do grafo G considere o parágrafo mostrado anteriormente nesta mesma seção. De acordo com o UMLS existem três sentidos possíveis para o termo *adjustment*. São eles: *individual adjustment*, *adjustment action* e *psychological adjustment*. Então, três grafos podem ser gerados a partir dos termos presentes no parágrafo até cada um dos possíveis sentidos. Um exemplo é o grafo do sentido *psychological adjustment*, apresentado na Figura 3.1, onde estão expressas CUIs de cada termo presente no parágrafo de acordo com o UMLS. Na Figura estão representados nas elipses escuras os termos encontrados no contexto. O retângulo cinza representa o termo candidato a desambiguação (CUI C0683269). Os demais são termos que estabelecem o relacionamento entre o termo candidato e os encontrados no contexto. Outros dois grafos representam os relacionamentos dos outros termos candidatos para desambiguação.

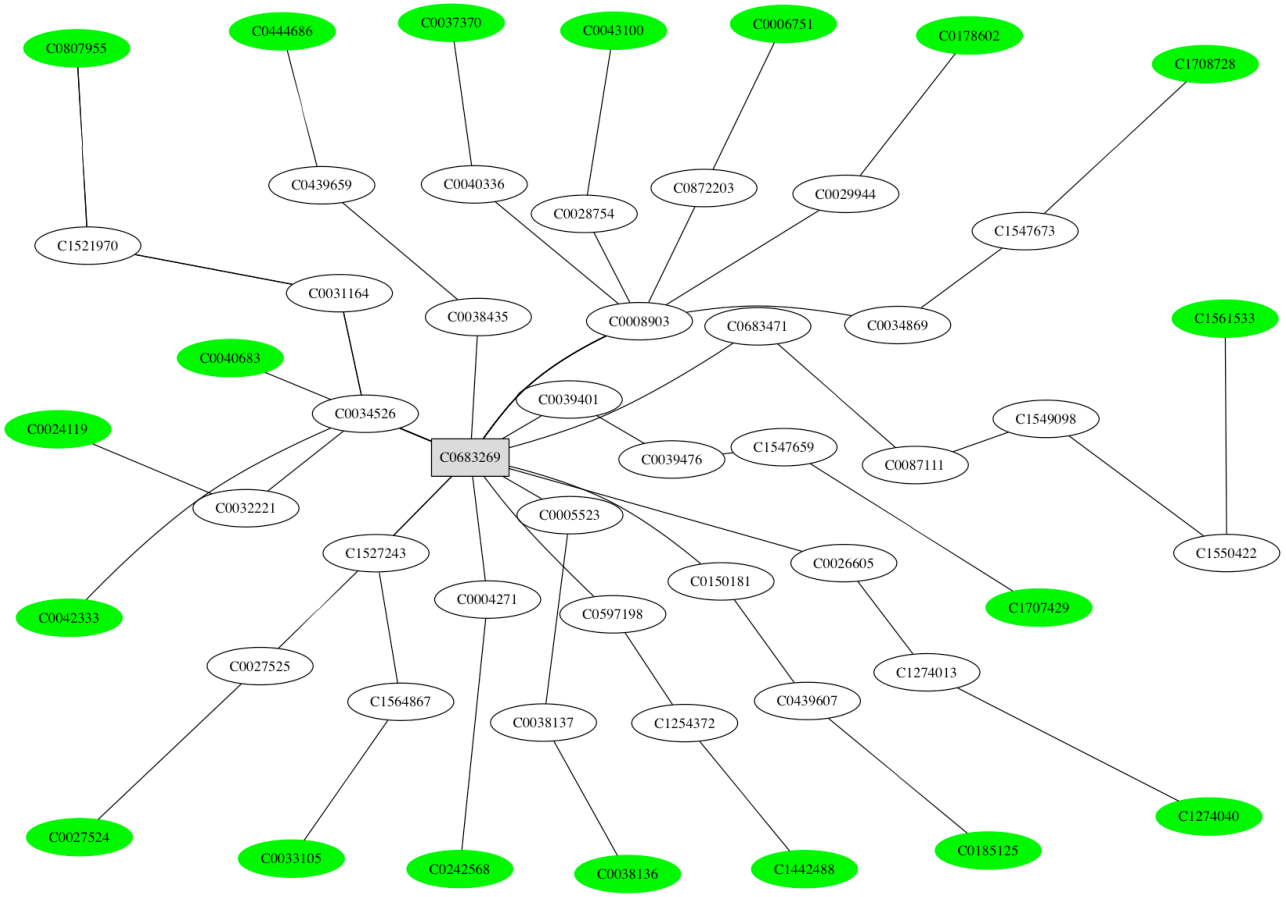

Figura 3.1: Grafo contendo os relacionamentos de CUIs para o termo *psychological adjustment*

3.2. Métricas de Conectividade

Para selecionar o sentido correto é necessária a classificação de cada vértice de acordo com a sua importância, baseado em alguma métrica de conectividade. Existem várias propostas dentre as quais foram selecionadas as mais discutidas e que obtiveram melhores resultados de acordo com Navigli e Lapata [36] e Navigli e Lapata [35]. São elas o Degree Centrality, Key Player Problem (KPP), PageRank, PageRank personalizado e Betweenness Centrality.

Degree Centrality, ou simplesmente *Degree*, é a maneira mais simples de medir a importância de um vértice. Ela é determinada pelo seu grau, ou seja, o número de arestas do vértice. Para isso temos:

$$\deg(v) = \left|\{\{u,v\} \in E : u \in V\}\right| \tag{3.1}$$

Um vértice é central se e somente se ele possui um alto grau. Da mesma forma, um vértice não conectado tem grau igual a zero. O grau de centralidade é o grau de um vértice normalizado pelo seu máximo grau, ou seja, o número de vértices do grafo com exceção de si mesmo. Temos então:

$$C_D(v) = \frac{\deg(v)}{|V|-1} \tag{3.2}$$

De acordo com o grafo da Figura 3.1 a importância do termo *Psychological Adjustment* (C0683269) é

$$C_D(C0683269) = \frac{11}{51} \text{ , ou } 0{,}21568627$$

Os demais termos candidatos, *Individual Adjustment* (C0376209) e *Adjustment Action* (C0456081) têm graus, respectivamente:

$$C_D(\text{C0376209}) = \frac{5}{52} \text{ , ou } 0{,}09615385 \text{ e } C_D(\text{C0456081}) = \frac{9}{51} \text{ , ou } 0{,}17647059.$$

Portanto, o mais alto grau encontrado é do termo C0683269 (*Psychological Adjustment*), o escolhido como sentido.

Com o **Key Player Problem** (KPP), um vértice é considerado importante se e somente se ele está relativamente próximo de todos os outros vértices [7]. Temos então:

$$KPP(v) = \frac{\sum_{u \in V, u \neq v} \frac{1}{dis(u,v)}}{|V| - 1} \tag{3.3}$$

onde o numerador é a soma das distâncias inversas entre v e todos os outros nodos. O denominador é o número de nodos do grafo, excluindo *v*. O KPP de um nodo desconectado é uma constante pequena, dada por $\frac{1}{K} = \frac{1}{|V|}$.

Por exemplo, considerando a Figura 3.1, temos o KPP(C06883269)=0,319149, o KPP(C0376209)=0,230769 e o KPP(C0456081)=0,277778. Portanto, C06883269 é escolhido como a melhor alternativa.

O algoritmo **PageRank** [8, 42] é um método para classificação de vértices de um grafo de acordo com a sua importância estrutural relativa. Ele foi originalmente desenvolvido para a classificação de páginas na Internet com base no número de páginas que contêm *links* para as mesmas. Nesta tese o algoritmo é descrito como um algoritmo para grafos genéricos.

O PageRank utiliza o modelo de *random walk*, onde um *random surfer* começa a percorrer o grafo a partir de um nodo arbitrário e, a cada passo, escolhe uma aresta de saída para um nodo qualquer e assim continua a visita de novos nodos. O *surfer* pode também decidir quando parar de seguir pelas arestas e se transferir para outro nodo no grafo. O PageRank de um vértice produz a probabilidade de um *random surfer* ser encontrado naquele vértice, assumindo que o movimento no grafo continua indefinidamente.

Especificamente, tomemos *G* como um grafo com *N* vértices (v_1, ..., v_n). Para um dado vértice v_i considere In(v_i) o conjunto dos vértices que apontam para v_i e d_j o número de arestas de saída do vértice v_j. O PageRank de um vértice v_i é definido na Equação (3.4).

$$P(v_i) = (1-c)\frac{1}{N} + c \sum_{v_j \in \mathrm{In}(v_i)} \frac{P(v_j)}{d_j} \tag{3.4}$$

O PageRank para um vértice v_i é a soma de dois termos. O coeficiente c, chamado *damping factor*, é um valor escalar entre 0 e 1. Ele modela a importância relativa de cada um dos dois termos da soma. O primeiro termo representa a probabilidade do *surfer* aleatoriamente saltar para qualquer nodo com igual probabilidade. O primeiro termo pode também ser visto como um fator de suavização (*smoothing factor*) tornando qualquer grafo aperiódico e irredutível, e assim garante que o cálculo do PageRank convirja para uma *unique stationary distribution*. O segundo termo modela a probabilidade de um *random surfer* chegar até v_i pelas arestas de um vértice v_j até o vértice v_i, dada pela soma das probabilidades de cada vértice v_j que tenha uma aresta para v_i vezes o peso de cada aresta, dada pelo inverso do grau de v_i.

PageRank é calculado pela execução iterativa do algoritmo da Equação 3.4 até a convergência abaixo de um determinado limiar ser atingida ou até um número pré-estabelecido de iterações terem sido executadas. O *damping factor* costuma ser configurado no intervalo [0,85..0,95]. Experimentos de Aguirre e Soroa [2], por exemplo, utilizaram um *damping factor* de 0,85.

A Figura 3.2 apresenta um grafo (a) e os valores de PageRank para esse grafo (b). Inicialmente, *P* para todos os nodos é inicializado com uma distribuição uniforme, i.e. 0,25. Com um fator de amortecimento de 0,85, na primeira iteração os valores de PageRank são atualizados da seguinte maneira:

$P(A^1)=$	$0{,}85 \times P(D^0) \times 1{,}0$	$+0{,}15 \times 0{,}25 = 0{,}25$
$P(B^1)=$	$0{,}85 \times P(A^0) \times 0{,}5$	$+0{,}15 \times 0{,}25 = 0{,}14$
$P(C^1)=$	$(P(A^0) \times 0{,}5 \times P(B^0) \times 1{,}0)$	$+0{,}15 \times 0{,}25 = 0{,}36$
$P(D^1)=$	$P(C^0) \times 1{,}0$	$+0{,}15 \times 0{,}25 = 0{,}25$

Os subscritos correspondem à iteração atual, i.e. $P(A_0)$ corresponde ao valor inicial e $P(A_1)$ corresponde à primeira iteração. A segunda iteração pode calcular $P(A_2)$ com base em $P(A_1)$, e assim por diante. Após algumas iterações a convergência é atingida e o valores do PageRank, apresentados no grafo (Figura 3.2.b) são obtidos.

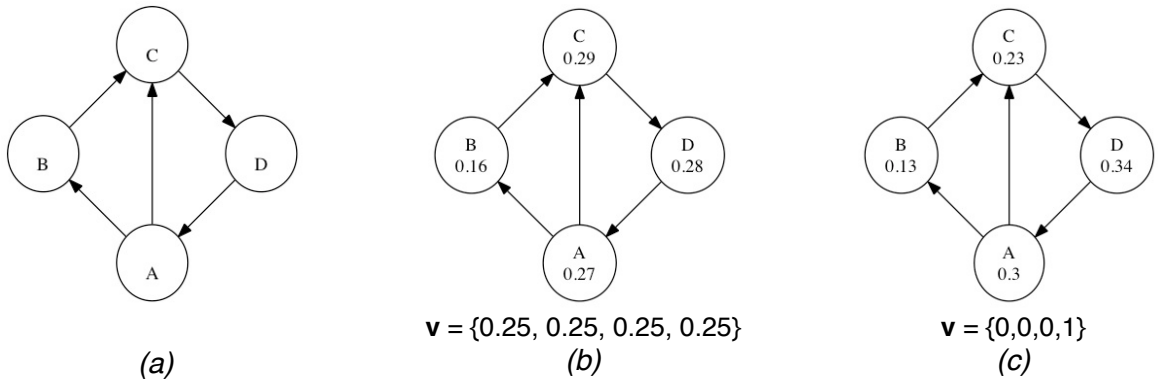

Figura 3.2: Exemplo de grafo empregando PageRank e Betweeness Centrality [3]

Em certas situações, incluindo WSD baseado em grafos, é desejável incluir informações sobre a importância relativa dos vértices no grafo. Dessa forma, dado um conjunto de vértices de interesse, é possível identificar quais outros vértices estão mais relacionados com eles no grafo. Por exemplo, considere o interesse em identificar quais nodos no grafo (Figura 3.2.a) estão mais relacionados com o nodo D (como apresentado na Figura 3.2.c)

Uma variação do algoritmo PageRank empregada no WSD é o **PageRank personalizado** [20]. Ela calcula a importância estrutural dos vértices de um grafo quando alguns deles são mais relevantes do que outros para uma determinada situação. Com o objetivo de apresentar o PageRank personalizado a Equação (3.5), proposta por Agirre *et al.*[3], é re-escrita numa forma compactada utilizando matrizes, como é apresentado a seguir. Considere M uma matriz de probabilidade de transição N x N, onde $M_{ji} = \frac{1}{d_i}$ se um caminho de um vértice v_i para v_j existe, senão é zero. Considere que v é um vetor estocástico normalizado N x 1 cujos valores são todos $\frac{1}{N}$. Então, o cálculo do Vetor PageRank P sobre o grafo G é equivalente à resolução da seguinte equação:

$$\mathbf{P}=c\mathit{M}\mathbf{P}+(1-\mathit{c})\mathbf{v} \tag{3.5}$$

No PageRank o vetor v é distribuído uniformemente, assim determinando probabilidades iguais para todos os vértices no grafo quando saltos aleatórios são feitos. No entanto, no PageRank personalizado o vetor v pode ser não uniforme e determinar probabilidades mais altas para determinados vértices, predispondo o resultado do vetor de PageRank para vértices preferenciais. Por exemplo, se toda a probabilidade for concentrada num único vértice v_x, todos os saltos aleatórios no percurso retornarão para v_x e consequentemente sua classificação será alta; além disso, a alta classificação de v_x fará com que todos os vértices da sua vizinhança tenham uma classificação alta também. A importância do vértice v_x na distribuição inicial de v se espalha pelo grafo durante as sucessivas iterações do algoritmo. Nesse caso, o vetor personalizado P representa a importância de cada vértice no grafo, relativa a v_x. O PageRank personalizado pode então ser calculado da mesma maneira que o PageRank tradicional.

A Figura 3.2 apresenta os resultados do PageRank em (b), onde v é uniforme, e o resultado do PageRank personalizado em (c) para o caso onde v recebe 0 para todos os vértices exceto D, que recebe 1. Para o PageRank tradicional, o vértice C recebe o valor de 0,29 (grafo da Figura 3.2.b), o que significa que um *random surfer* nesse grafo pode gastar 29% do seu tempo nesse nodo [2]. C tem a maior classificação entre os nodos, e por essa razão o nodo C é o mais importante nodo no grafo. Por outro lado, se for utilizado o PageRank personalizado e o *random surfer* fizer todos os saltos aleatórios para o nodo D (grafo da Figura 3.2.c), então a classificação de D é a mais alta, seguida pelo nodo A, que está conectado diretamente a D, e os nodos C e B.

O algoritmo **Betweenness Centrality** calcula para um dado vértice v a fração dos menores caminhos entre dois vértices que passam por v [16]. Formalmente, o Betweenness Centrality é definido como:

$$betweenness(v) = \sum_{s,t \in V : s \neq v \neq t} \frac{m_{st}(v)}{m_{st}} \tag{3.6}$$

onde m_{st} é o número de menores caminhos de s para t, e $m_{st}(v)$ é o número de caminhos de s para t que passam pelo vértice v. A normalização é feita pela divisão do *betweeness*(v) pelo número máximo de pares de vértices excluindo v, s e t. Temos então:

$$C_B(v) = \frac{betweenness(v)}{(|V|-1)(|V|-2)} \tag{3.7}$$

O objetivo desta métrica é identificar os vértices que estão envolvidos com o maior número de menores caminhos entre outros dois vértices, em comparação com o número total de pares de vértices. O betweenness centrality de um nodo desconectado é zero.

3.3. Avaliação de Métricas

Navigli e Lapata [35] desenvolveram um estudo experimental para avaliar oito métricas de conectividade utilizando três *corpora*, o *corpus* SemCor [32], o *corpus* Senseval-3 [49], e o *corpus* Semeval-2007 [44]. O objetivo era comparar o desempenho de cada métrica com as demais, além de dois *baselines*. Navigli e Lapata utilizaram a WordNet como fonte de conhecimento (grafo) para distinguir os sentidos, assim como as relações lexicais e semânticas dos *corpora*.

O *corpus* **SemCor** é composto por 352 documentos. Destes, 186 documentos têm substantivos, verbos, adjetivos e advérbios identificados e anotados com seus respectivos sentidos. Os 166 documentos restantes tiveram apenas os verbos anotados com seus sentidos. O *corpus* foi criado para viabilizar exemplos de sentidos em seus contextos. Uma curiosidade é a de que a ordem de importância de cada sentido na WordNet é baseada na sua frequência no SemCor. O SemCor é utilizado em experimentos supervisionados e não supervisionados em WSD. O **Senseval-3** e o **Semeval-2007** são subconjuntos do *corpus* Wall Street Journal [43]. Eles contêm, respectivamente, 3.037 e 465 palavras anotadas com sentidos da WordNet. Cada um dos três *corpora* utiliza uma versão diferente da WordNet. SemCor utiliza a versão 1.6, Senseval-3 utiliza a versão 1.7.1 e o Semeval-2007 a versão 2.1. Para a avaliação conjunta a anotação da WordNet nos três *corpora* a anotação foi normalizada para a WordNet 2.0, utilizando mapeamentos de sentidos à disposição em http://nlp.lsi.upc.edu/ (Último acesso em 21 de Maio de 2013).

Além da versão disponível publicamente, uma versão estendida criada por Navigli, chamada EnWordNet [33], também foi utilizada. O grafo deste léxico contém cerca de 60.000 arestas adicionais que relacionam conceitos por intermédio de relações sintáticas chamadas *collocations*. Essa informação não está codificada explicitamente na WordNet. A EnWordNet foi elaborada a partir da lista de expressões presentes, principalmente, nos dicionários Oxford Collocations [12] e no Logman Language Activator [50]. As expressões representam pares consistindo num elemento base (e.g. o verbo *drink*) e seu *collocate* (e.g. o substantivo *water*), desde que presentes na WordNet. Após um processo de desambiguação semi-automático do sentido na WordNet as novas relações foram reunidas e acrescentadas ao conjunto original da WordNet.

Dois *baselines* foram utilizados. O primeiro é baseado na seleção aleatória de um sentido e o segundo utiliza o algoritmo de Lesk para o WSD [27]. O trabalho se propõe a desambiguar uma palavra com a comparação das palavras encontradas em definições de um dicionário (textos com informações adicionais presentes na WordNet) com as palavras encontradas no contexto da palavra ambígua. O sentido que obtivesse o maior número de coincidências seria então o escolhido.

Navigli e Lapata [35] implementaram um conjunto de experimentos para avaliar diferentes configurações de métricas, recursos estruturados e *corpora*. Os resultados do comparativo,

considerando os algoritmos para grafos apresentados anteriormente, estão reunidos na Tabela 3.1. Os algoritmos *Degree*, PageRank e *Betweeness* obtiveram os melhores resultados. Foram avaliadas as performances nas configurações *all-words* (todas as palavras do texto são avaliadas) e *lexical sample* (apenas as palavras consideradas polissêmicas são avaliadas). Os *corpora* utilizados nos experimentos são o SemCor, Senseval-3 e o Semeval-2007, anteriormente apresentados. Apenas a métrica com os melhores no SemCor foi avaliada nos demais *corpora*. Como recursos estruturados, para criação do grafo de sentidos, foram utilizadas a WordNet versão 2.0 e a versão estendida EnWordNet [38]. Em cada um dos experimentos foram mensuradas a precisão (*P*), *recall* (*R*) e medida F1. A precisão corresponde ao número de sentidos corretos no conjunto de sentidos retornados, o *recall* corresponde ao número de sentidos corretos identificados em razão do conjunto total de sentidos identificados no *corpus*, e F1 a combinação entre precisão e *recall* determinada por $\frac{2PR}{P+R}$.

Tabela 3.1: Resultados do comparativo de métricas, adaptado de [35] onde AW e LS significam *all-words* e *lexical sample*, respectivamente

	SemCor				Senseval-3	Semeval-2007
	WordNet		EnWordNet		EnWordNet	
	AW	LS	AW	LS	AW	AW
Degree	**50,01**	**37,80**	**56,62**	**46,03**	52,9	43,1
PageRank	49,76	37,49	56,46	45,83	-	-
Betweeness	48,72	36,20	56,48	45,85	-	-

Segundo Navigli e Lapata os resultados dos experimentos com as métricas *Degree* e PageRank no *corpus* SemCor, são estatisticamente similares. Independentemente da fonte de conhecimento externa utilizada o valor de um nodo para o PageRank é proporcional ao seu grau em grafos não-dirigidos. Por outro lado, uma diferença significativa entre eles é a da complexidade. *Degree* é considerado $O(n)$ e o PageRank $O(n^2)$, ou seja, o tempo para a análise dos termos cresce de forma linear e quadrática, respectivamente.

Outra constatação importante é a respeito das fontes de conhecimento externas. Relações semânticas mais densas, presentes na EnWordNet, aumentaram em até 9% a performance das métricas (quando na modalidade *lexical samples*). Este aumento está relacionado ao fato destas abordagens se beneficiarem do número de relacionamentos para distinguir melhor a importância dos vértices. O número médio de relacionamentos exclusivos da EnWordNet (*collocations*) presentes nos termos selecionados pela métrica *Degree* é de 20,5 arestas. A WordNet original tem como relacionamentos mais expressivos os de hiperonímia e hiponímia, representando juntos o número de 9,29 arestas em média nos termos selecionados corretamente pela métrica *Degree*. Navigli e Lapata consideram que, além de um grau maior, os relacionamentos exclusivos da EnWordNet estabelecem conexões transversais importantes e que não necessariamente fazem parte de uma taxonomia. Desta forma, há um indicativo de que relações do tipo *collocations* são importantes para o WSD.

Considerando o *corpora* Senseval-3 e Semeval-2007 a métrica de *Degree* não obteve melhores resultados que os campeões das duas competições (65,2 e 59,1 respectivamente). No

entanto, dentre os que utilizaram métodos não supervisionados a métrica obteve um resultado significativamente melhor que os demais concorrentes (45,8 e 40,2 respectivamente). Por fim, o uso desta métrica é relevante quando comparado com outros métodos mais sofisticados sem a necessidade de treino ou qualquer outra informação morfossintática. No entanto, o emprego de uma fonte de conhecimento externo com maior número de relacionamentos entre os termos, é determinante para o WSD.

4. WSD NO PLN EM BIOMEDICINA

O WSD é um problema explorado em diferentes áreas do conhecimento, do geral ao especializado. No domínio especializado o Biomédico é de grande destaque. Neste domínio específico, dois grupos de métodos de classificação são geralmente utilizados: os supervisionados e os não supervisionados, como foi apresentado no Capítulo 3. Além dos métodos, o *corpus* NLM-WSD [53] é utilizado com frequência para a comparação entre as diferentes propostas de solução para o problema. Dada a sua importância na fundamentação de conceitos deste trabalho, o *corpus* NLM-WSD é apresentado num primeiro momento. O capítulo segue com a apresentação das abordagens mais utilizadas.

4.1. O *Corpus* NLM-WSD

O *corpus* NLM-WSD foi construído a partir de 409.337 resumos (título + resumo) presentes na base MEDLINE em 1998. Os conceitos do metatesauro UMLS presentes nos resumos foram identificados pelo *parser* MetaMap [4]. O *parser* faz o mapeamento dos conceitos de acordo com uma versão do UMLS. É possível observar, a título de detalhe, que na ocasião da construção do *corpus* a versão utilizada foi a de 1999. O *parser* identificou 4.051.445 sintagmas ambíguos, ou seja, que foram relacionados pelo *parser* com duas ou mais entradas no UMLS. Eles são o equivalente a 11,7% dos mais de 34 milhões de sintagmas identificados nos 409.337 resumos processados.

Três tipos de ambiguidade foram identificadas de acordo com Weeber *et al.* [53]:

1. **Ambiguidade simples**: corresponde a conceitos que possuem a mesma grafia mas significados diferentes, por exemplo, o conceito *activity* que possui três entradas no UMLS. Esse tipo de ambiguidade representa 94,3% de todos os casos ambíguos identificados pelo MetaMap.
2. **Ambiguidade lexical**: envolve aqueles conceitos que podem corresponder a mais de uma entrada no UMLS se consideramos as suas flexões, número etc. Um exemplo é a palavra *reported* que o MetaMap associa a *reporting*, *reports* e *report*. Esse caso de ambiguidade representa 5,5% dos casos ambíguos identificados.
3. **Ambiguidade complexa**: é o tipo de ambiguidade (difícil e rara) onde um sintagma pode corresponder a diferentes entradas no UMLS. Por exemplo, *reproductive health policies* pode ser interpretado como "*reproductive health" policies*, *reproductive "health policies"* ou simplesmente *reproductive health policies* (cada palavra ou grupo de palavras com uma definição).

Na construção do *corpus* apenas as ambiguidades de tipo simples foram utilizadas. Foram então selecionados para a anotação manual os 50 conceitos ambíguos mais frequentes. Destes, alguns foram desconsiderados por não terem definições ou relacionamentos consistentes no UMLS. Estas eram condições importantes pois foram utilizadas pelos anotadores como fonte de informação na anotação. Para cada um dos conceitos frequentes selecionados foram coletadas aleatoriamente 100 instâncias (sintagmas) para anotação manual, num total de 5.000 instâncias. Os conceitos selecionados tinham entre 3 e 6 alternativas para desambiguação. No entanto, em 17 dos 50 conceitos, uma ou mais alternativas não foram utilizadas por terem sentidos muito próximos para uma distinção prática. Um exemplo é o conceito *depression,* com as alternativas *depression motion*,

depressive episode, e *mental depression*. Os dois últimos foram considerados muitos próximos em significado e portanto apenas o último, *mental depression*, foi utilizado. Além das alternativas presentes no UMLS, os anotadores poderiam classificar um conceito como *none* caso não houvesse uma resposta adequada.

Os resultados da anotação foram analisados com o emprego de até três métodos, para então selecionar a alternativa final. O primeiro consiste em selecionar a alternativa com a maioria dos votos. Caso não haja uma alternativa com uma diferença de dois votos ou mais para as demais, a estatística Kappa (*k*) [15] foi utilizada para excluir os resultados que divergem dos demais. Por fim, se a exclusão pelo Kappa não altera os resultados, o método *Latent Class Analysis* (LCA) [10] foi utilizado para definir a classificação. A Figura 4.1 relaciona os conceitos anotados no *corpus* NLM-WSD. Dos 50 conceitos selecionados, apenas 12 utilizaram o último método para desambiguação. Destes, 159 instâncias tiveram de ser discutidas pelo grupo de anotadores para a classificação final. Em Weeber *et al.* [53] é recomendada a utilização dos 38 conceitos que obtiveram maior concordância entre os anotadores, cerca de 3.800 instâncias. Na Figura 4.1 as palavras com um asterisco (*) representam os 12 conceitos mais difíceis para a anotação.

Adjustment*	Energy	Growth	Pathology	Single
Association	Evaluation*	Immunosuppression*	Pressure	Strains
Blood_pressure*	Extraction	Implantation	Radiation	Support*
Cold	Failure*	Inhibition	Reduction	Surgery
Condition*	Fat	Japanese	Repair	Transient
Culture	Fit	Lead	Resistance	Transport
Degree	Fluid	Man	Scale	Ultrasound
Depression	Frequency	Mole	Secretion	Variation*
Determination*	Ganglion	Mosaic*	Sensitivity*	Weight
Discharge	Glucose	Nutrition*	Sex	White

Figura 4.1: Lista de conceitos anotados no NLM-WSD,
com um asterisco (*) sinalizando os 12 casos mais complexos

O *corpus* na versão 3 contém um arquivo para cada conceito ambíguo anotado. Cada arquivo contém um conjunto de 100 resumos e em cada um deles uma instância anotada do conceito ambíguo, com os conceitos candidatos e o selecionado pelos anotadores. O conceito é anotado com a sua posição no texto além de sua CUI, que corresponde à versão 1999 do UMLS.

4.2. Abordagens Supervisionadas e Não-supervisionadas

Abordagens baseadas no aprendizado supervisionado são comuns no WSD de textos de Biomedicina, a exemplo daquelas de Joshi [24], Liu [28] e Savova [47]. Contudo, elas exigem

exemplos etiquetados para o treinamento que podem não estar à disposição, ou representar alto custo para serem elaborados. Essa limitação significa que as abordagens supervisionadas podem desambiguar uma pequena amostra de palavras para a qual um conjunto de dados de treino foi elaborado, e isso limita sua utilização na prática.

Abordagens não-supervisionadas, por sua vez, não necessitam de exemplos etiquetados. Por fazerem uso de recursos estruturados como fonte de conhecimento externa, não há necessidade de um conjunto para treino e teste. Três exemplos são relevantes neste aspecto e utilizam o *corpus* NLM-WSD [53] como conjunto de teste.

McInnes [30] utiliza o UMLS como fonte de conhecimento sobre os conceitos candidatos a desambiguação. As palavras do contexto do conceito ambíguo são comparadas com as palavras presentes em definições, ou comentários dos conceitos candidatos no UMLS. Na ausência destes, são utilizadas as definições dos tipos semânticos (*Semantic Types* - ST do UMLS) relacionados aos conceitos. Uma medida de distância entre os candidatos e o contexto identifica qual conceito é o mais apropriado. Essa abordagem foi avaliada com 13 conceitos do *corpus* NLM-WSD e atingiu uma performance de 48,11% de acerto. Outro exemplo é o de Humphrey [21], que propõe o *ranking* dos STs relacionados aos conceitos candidatos. Essa classificação é obtida com base numa medida de relacionamento das palavras do contexto e o ST de cada candidato. Contudo, essa proposta é semi-supervisionada. O peso do relacionamento entre palavra e ST é calculado por um processo que utiliza 4.000 textos da MEDLINE e as palavras contidas nos conceitos associados a cada ST para estabelecer o *ranking* entre palavra e ST. Além disso, o resultado não é conclusivo quando os conceitos candidatos estão associados a um mesmo ST, o que lhes confere o mesmo peso.

Outro tipo de abordagem, dentre as não-supervisionadas, é a que faz uso da estrutura de um grafo como fonte de conhecimento externo. Os meios para modelar e empregar essa abordagem foram apresentados no Capítulo 3. Os trabalhos que empregam esse tipo de abordagem, no WSD independente de domínio, utilizam com frequência a WordNet como fonte de conhecimento estruturado (grafo). Como *corpus* para teste, uma ou mais versões do *corpus* Semeval costumam ser utilizadas [36, 48, 2, 51]. Do mesmo modo, métodos baseados em grafos foram empregados em domínios específicos, com é o caso do domínio de Biomedicina. Neste cenário o trabalho de Agirre *et al.* [3] é um exemplo e motivação desta proposta.

4.3. Abordagem de Grafos

Agirre *et al.* [3] propõem a utilização da abordagem baseada em grafos para o domínio da Biomedicina. Nesse trabalho, o algoritmo de PageRank personalizado (apresentado na Seção 3.2) é empregado no WSD, com o uso do metatesauro UMLS como fonte de conhecimento. Os relacionamentos presentes no UMLS são utilizados na construção de um grafo, que é então analisado pelo algoritmo. Assim, o *ranking* de cada conceito candidato é gerado com base na importância relativa do mesmo em relação aos demais conceitos do contexto do conceito ambíguo. Esse algoritmo foi utilizado anteriormente num cenário independente de domínio, utilizando a WordNet como base de conhecimento. Ele resultou em melhores resultados que as outras propostas baseadas em grafos [2].

Utilizando o *corpus* NLM-WSD de Weeber *et al.* [53], Agirre *et al.* comparam os resultados do algoritmo PageRank personalizado com a performance de dois *baselines*. Além disso, os experimentos foram comparados com os resultados de McInnes [29], que utilizou um subconjunto do *corpus* NLM-WSD (cerca de 58% deles são casos "difíceis" da Figura 4.1). O software elaborado para os experimentos utiliza três fontes como entrada (Figura 4.2). A **primeira entrada** é um dicionário composto por todos os conceitos do UMLS mapeados no *corpus*, incluindo as

palavras e seus CUIs. A **segunda entrada** são os contextos de cada instância de conceito ambíguo presente no *corpus* NLM-WSD. Nela estão relacionados os conceitos mapeados no *corpus* em uma janela de 20 conceitos (10 antes e 10 após o conceito ambíguo). A **terceira entrada** são os relacionamentos entre os conceitos presentes no UMLS. A versão do UMLS utilizada nos experimentos é a 2007AB.

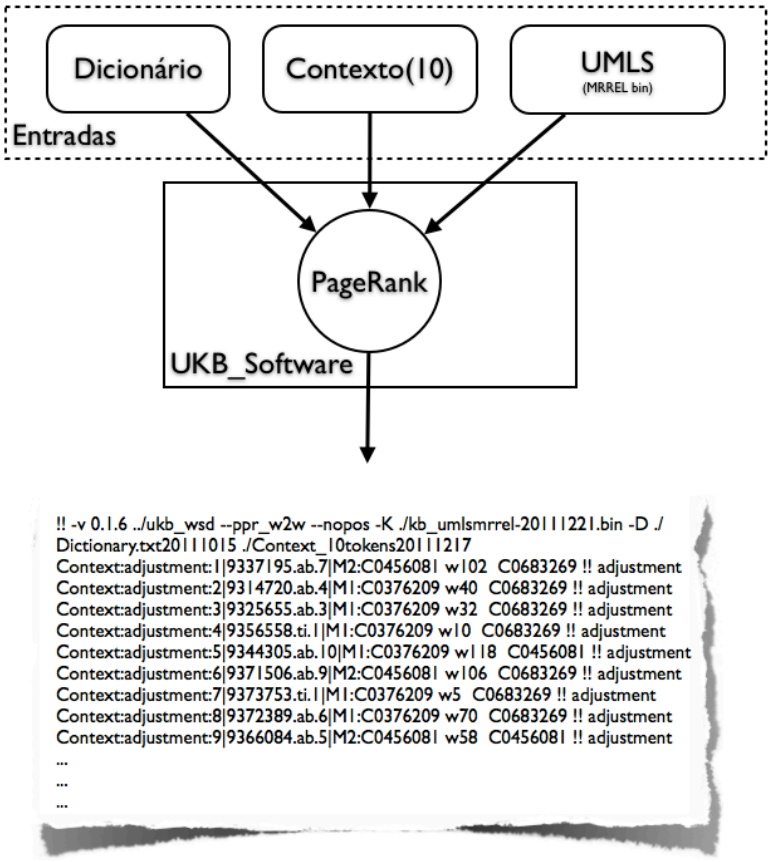

Figura 4.2: *Overview* do experimento de Agirre *et al.* [3]

As três entradas são resultado do pré-processamento de diferentes fontes de informação por intermédio de *scripts*, que extraem ou relacionam as informações das mesmas. O dicionário e o contexto foram elaborados a partir do *corpus* NLM-WSD. Agirre *et al.* [3] utilizam a versão 3 do *corpus* NLM-WSD. O pré-processamento executa as seguintes etapas:

1) anotar os demais conceitos presentes em cada resumo utilizando o *parser* MetaMap,
2) construir um Dicionário de entrada com todos os conceitos encontrados e seus respectivos CUIs, e
3) extrair os conceitos do contexto de acordo com a janela estabelecida.

MetaMap [4] é um *parser* que relaciona os conceitos de Biomedicina presentes em um texto com conceitos do metatesauro UMLS. Ele é o mesmo *parser* utilizado na anotação do NLM-WSD, embora em Agirre *et al.* [3] uma versão mais recente tenha sido utilizada (2007AB). O algoritmo que implementa seu funcionamento executa cinco passos: o *parsing*, a geração de variantes, a recuperação de candidatos, a avaliação de candidatos e a construção do mapeamento.

A etapa de ***parsing*** faz prioritariamente a identificação de sintagmas nominais. O objetivo é reduzir o escopo de possibilidades e consequentemente reduzir o processamento. A identificação

dos sintagmas tem como base o léxico SPECIALIST [9], que é parte do UMLS. Além disso, também são identificadas as categorias morfossintáticas das palavras, presentes nos sintagmas e que não representem *stop phrases* (Figura A.1 do Anexo A). A etapa de **geração de variantes** utiliza, além de uma base de dados suplementar do autor, o conhecimento presente no léxico SPECIALIST. As variações consistem numa relação entre cada palavra do sintagma e seus acrônimos, abreviações, etc. Por exemplo, considerando a palavra *ocular*, temos seus sinônimos, flexões e derivações apresentadas na Figura 4.3. A hierarquia representa a ordem em que elas foram criadas. Para cada variação é identificada sua categoria morfossintática e uma pontuação da distância em relação à palavra original. Flexões (f) com peso 1. Sinônimos (s) ou acrônimos e suas expansões com peso 2. Finalmente, derivações (d) com peso 3. A palavra *ophthalmia* é um substantivo cuja pontuação é 7, por ser a derivação de um sinônimo (*ophthalmic*) do sinônimo (*eye*) de *ocular*.

```
ocular {[adj], 0=""}
        eye {[substantivo], 2="s"}
                eyes {[noun], 3="sf"}
                optic {[adj], 4="ss"}
                ophthalmic {[], 4="ss"}
                        ophthalmia {[substantivo], 7="ssd"}
        oculus {[substantivo], 3="d"}
                oculi {[substantivo], 4="df"}
```

Figura 4.3: As variações de *ocular*, adaptado de [4]

A etapa de **recuperação de candidatos** relaciona todas as entradas que contêm pelo menos uma das variantes de uma palavra no UMLS. Isso significa que um conceito composto por mais de uma palavra, mas que contém uma das variantes, é relacionado como candidato. Com todas as entradas identificadas, a etapa de **avaliação de candidatos** é executada. As palavras do sintagma são avaliadas em relação a cada candidato a conceito do UMLS, de acordo com o peso médio de quatro métricas: centralidade, que mede o envolvimento com o núcleo do sintagma; variação, o envolvimento com a pontuação da distância das variações; cobertura e coesão, onde é medido o quanto um candidato combina com o texto do sintagma, e em quantas palavras. Os nove candidatos para o sintagma *ocular complications* são apresentados na Figura 4.4.

```
861 complications <1> (Complication)
861 complications <3> (Complications Specific to Antepartum or
Postpartum)
777 Complicated
694 Ocular
638 Eye
838 Eye NEC
611 Ophthalmic
611 Optic (Optics)
588 Ophthalmia (Endophthalmitis)
```

Figura 4.4: As variações de *ocular complications*, adaptado de [4]

Para demonstrar o processo completo e o resultado final, considere o conceito *cold* na seguinte frase extraída do *corpus* NLM-WSD (Figura 4.5):

> *… use. OBJECTIVE: To evaluate antibiotic-prescribing practices for children younger than 18 years who had received a diagnosis of* ***cold****, upper respiratory tract infection (URI), or bronchitis in the United States. DESIGN: Representative national survey of practicing physicians participating in the National Ambulatory Medical Care Survey conducted in 1992. ...*

Figura 4.5: Trecho de um resumo contendo o conceito *cold*

Após o pré-processamento de anotação dos conceitos com o *parser* MetaMap, e extração do contexto, temos o resultado exposto na Figura 4.6. Cada conceito anotado é seguido de sua classe gramatical, posição no resumo e uma indicação de se é ou não o conceito ambíguo do resumo. Por exemplo, o conceito *diagnosis#n#w40#0* é um substantivo (#n) na posição 40 (#w40) e não corresponde ao conceito ambíguo do resumo (#0). Já o conceito *cold#n#w41#1* é outro substantivo, na posição 41 e corresponde ao conceito ambíguo (#1).

> *Context:cold:45|9516004.ab.3|M2:C0009443*
> *use#n#w31#0 evaluate#n#w32#0 antibiotic#n#w33#0*
> *prescribing#n#w34#0 practice#n#w35#0 children#n#w36#0*
> *young#n#w37#0 years#n#w38#0 received#n#w39#0* ***diagnosis#n#w40#0***
> ***cold#n#w41#1*** *upper_respiratory_tract_infection#n#w42#0*
> *bronchitis#n#w43#0 united_states#n#w44#0 representative#n#w45#0*
> *national_survey#n#w46#0 practice#n#w47#0 physicians#n#w48#0*
> *participating#n#w49#0*
> *national_ambulatory_medical_care_survey#n#w50#0 conduct#n#w51#0*

Figura 4.6: O conceito *cold*

Para utilizar o PageRank personalizado no WSD, o UMLS é pré-processado para que sejam extraídos os relacionamentos entre CUIs na forma de um grafo. Na versão UMLS utilizada em Agirre *et al.* [3] as seguintes CUIs estão associadas ao conceito ‘cold’:

C0009443: *‘Common Cold*’,
C0009264: ‘*Cold Temperature*’ e
C0234192: ‘*Cold Sensation*’.

O Metatesauro contém informações sobre os relacionamentos entre CUIs na forma de bases de dados em tabelas. A tabela MRREL reúne diferentes tipos de relacionamentos entre CUIs. Ela também relaciona a fonte de onde a relação foi obtida. Como foi apresentado na Seção 2.2., a mesma relação pode ser encontrada em múltiplas fontes. Por exemplo, as CUIs C0009443 e C0035243 são encontradas em quatro fontes diferentes.

Além das relações entre CUIs presentes na tabela MRREL, as relações de coocorrência entre CUIs são encontradas na tabela MRCOC. A tabela MRCOC inclui detalhes sobre o peso da relação de coocorrência entre conceitos, baseada no número de coocorrências identificada [40]. Contudo, os resultados encontrados em Agirre *et al.* [3] demonstram que as relações de coocorrência não levaram a melhores resultados. A hipótese levantada por aqueles autores é de que as relações de coocorrência alteram negativamente a topologia do grafo, prejudicando a performance do algoritmos de PageRank personalizado.

A conversão da informação contida nas tabelas em um grafo é simples. Os conceitos se tornam vértices e as relações presentes nas tabelas se tornam as arestas entre eles. Nenhum peso é associado às relações que são extraídas da tabela MRREL. Por outro lado, a tabela MRCOC pode produzir subgrafos com o emprego dos pesos das relações de coocorrência. Considere a tabela MRREL, que obteve os melhores resultados, e o conceito candidato C0009443: *'Common Cold*'. Um grafo pode ser criado de acordo com método apresentado na Seção 3.1. Considerando o contexto apresentado na Figura 4.6 temos o grafo da Figura 4.7.

Com as três entradas estabelecidas o experimento conduzido por Agirre *et al.* empregou três configurações diferentes com o algoritmo de PageRank personalizado. Duas variam a quantidade de tabelas do UMLS utilizadas como fonte de conhecimento: apenas a MRREL ou a MRREL e a MRCOC. Quando empregada, a tabela MRCOC adiciona relacionamentos ao grafo com base nas relações com maior probabilidade de ocorrência. Isso significa que dados sobre a frequência das coocorrências não são utilizados pelo algoritmo de PageRank como um peso de uma relação. A terceira configuração utiliza um subconjunto do *corpus* NLM-WSD [29]. Além destas configurações o algoritmo foi comparado com uma versão estática e outra aleatória. A versão estática utiliza todo o grafo UMLS para selecionar o sentido, sem considerar o contexto. A versão aleatória simplesmente seleciona um sentido de forma aleatória. A Tabela 4.1 resume os resultados dos experimentos de Agirre *et al.* [3].

Os dois primeiros valores representam os resultados do PageRank personalizado com as duas configurações de base de conhecimento. O método estático resultou numa performance menor em relação ao algoritmo proposto em [3], assim como o método aleatório. Em comparação ao trabalho de McInnes, o algoritmo de PageRank obteve uma performance 6,9% melhor. Uma análise parcial dos resultados, palavra por palavra, também é feita por Agirre *et al.*, onde os resultados percentuais são apresentados para cada conceito analisado. A Tabela 5.1 apresenta essas informações em conjunto com resultados de experimentos implementados nesta proposta de tese, abordados no Capítulo 5.

A relevância dos resultados obtidos por Agirre *et al.* [3] é uma das discussões apontadas no trabalho. A Tabela 4.2 apresenta um comparativo entre os resultados de Humphrey *et al.* [21], McInnes [29] e Agirre *et al.* [3]. Nela estão relacionados 13 dos 50 conceitos presentes no NLM-WSD. Esse conjunto reduzido foi inicialmente estabelecido por Humphrey *et al.* [21], e foi então utilizado em McInnes [29]. Humphrey *et al.* obtiveram a maior parte dos melhores resultados individuais. Cerca de 76% dos conceitos obtiveram o melhor resultado. A média atingida foi de 68,26% de acerto. A destacar que este experimento relatado em [21] se restringe a um subconjunto do *corpus* NLM-WSD. Além disso, ele não emprega a abordagem de grafos que foi selecionada para nossa proposta. Como veremos mais adiante neste trabalho, a média de acerto de Humprhrey não foi usada como *baseline* para nossa pesquisa, por este motivo.

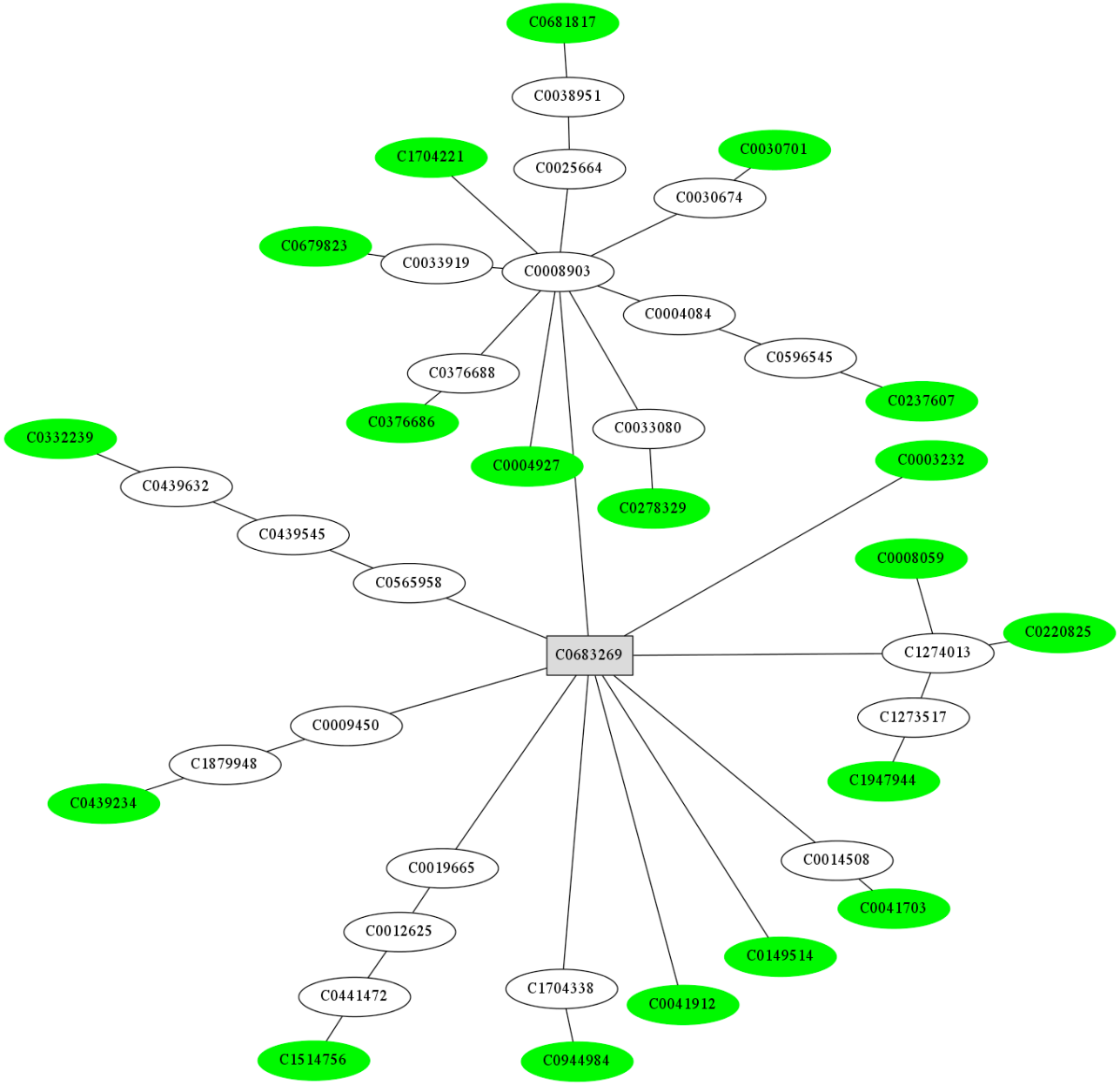

Figura 4.7: Grafo do conceito C0009443: *'Common Cold'*

A abordagem de McInnes não obteve melhores resultados individuais e resultou numa média de 48,11% de acerto. A abordagem de Agirre *et al.* atingiu uma média de 56,14% de acerto. Obteve os melhores resultados individuais em 3 conceitos (cerca de 23% do total). Em outros dois casos ficou próxima dos melhores resultados. O fato a ser destacado é que a abordagem de Humphrey *et al.* utiliza um método semisupervisionado. Em comparação aos demais métodos não supervisionados, a abordagem de grafos aumenta consideravelmente a performance geral de um sistema de WSD neste conjunto reduzido de conceitos para teste.

Sanderson [46] indica que uma taxa mínima de 90% de acerto é necessária para que um sistema de WSD seja útil para a tarefa de Recuperação de Informação (RI), donde salientamos o espaço de melhoria em relação aos resultados atuais. Em razão de as taxas de acerto, geralmente, não atingirem este patamar, a afirmação de Sanderson ainda não foi descartada. No entanto, mesmo com taxas de acerto menores, um sistema de WSD melhora a performance de sistemas de RI [46].

Tabela 4.1: Resultados de Agirre *et al.* [3]

Método	**Base de Conhecimento**	**Taxa de acerto (%)**
NLM-WSD completo		
PageRank personalizado	MRREL	**68,1**
PageRank personalizado	MRREL + MRCOC	65.5
Estático	MRREL	58,4
Aleatório	--	45,6
NLM-WSD de McInnes [29]		
PageRank personalizado	MRREL	**55,0**
McInnes [29]	--	48,1

Tabela 4.2: Comparativo de resultados

Conceito	**Humphrey *et al.* [21]**	**McInnes [29]**	**Agirre *et al.* [3]**
Adjustment	**76,67**	44,57	35,50
Blood_pressure	41,79	38,38	**48,00**
Degree	**97,73**	70,31	93,80
Evaluation	**59,70**	51,52	50,00
Growth	**70,15**	63,64	37,00
Immunosuppression	74,63	50,51	62,00
Mosaic	**67,69**	37,50	66,00
Nutrition	**35,48**	25,00	32,60
Radiation	**78,79**	57,73	53,10
Repair	**86,36**	37,31	76,50
Scale	60,47	51,56	**84,60**
Sensitivity	**82,86**	48,00	27,50
White	55,00	49,44	**63,30**
Média	**68,26**	48,11	56,14
Média dos difíceis	**62,68**	42,21	45,94

Outro aspecto é a dificuldade em desambiguar os conceitos do *corpus* NLM-WSD, uma vez que a concordância entre os anotadores obteve um Kappa 0,47 [47]. Dentre os 12 conceitos considerados difíceis, segundo Weeber *et al.* [53], 7 fazem parte do conjunto presente na Tabela 4.2 (sublinhados na coluna Conceito). Considerando apenas os conceitos difíceis, a abordagem de Humphrey *et al.* também obteve os melhores resultados. A comparação entre a média geral e a média dos difíceis demonstra que a abordagem de Agirre *et al*. teve uma perda de 18,17% na taxa de acertos (de 56,14% para 45,94%). Enquanto isso, Humphey *et al.* e McInnes tiveram uma perda de 8,18% (de 48,11% para 42,21%) e 12,27% (de 68,26% para 62,68%), respectivamente. Portanto, se observa que a abordagem de Agirre *et al.* é a que se beneficia mais dos casos de ambiguidade simples.

O estudo destas propostas para o WSD não supervisionado e, em especial, a abordagem baseada em grafos, levou à investigação de outros algoritmos que pudessem ser aplicados a este cenário. O próximo capítulo apresenta novas alternativas de algoritmos para o domínio da Biomedicina e uma nova proposta de abordagem utilizando grafos.

5. MODELO SIMPLES: COMPARATIVO ENTRE MÉTRICAS

A tarefa de desambiguar conceitos de Biomedicina por intermédio de abordagens baseadas em grafos é a principal motivação desta pesquisa. Identificar métodos que conduzam a novos resultados exige procedimentos de proposta, implementação, teste e avaliação de métricas. Os capítulos 3 e 4 apresentam modelos simples para a seleção do sentido de uma palavra ambígua. Em outras palavras, cada proposta emprega apenas uma métrica como método. Por esta razão, de forma geral, três propostas foram selecionadas para estabelecer um comparativo. Dentre elas, o trabalho de Agirre *et al.* [3] é a principal referência para experimentação e fundamentação desta tese.

Agirre *et al.* [3], em trabalho apresentado no Capítulo 4, propõem a experimentação do algoritmo de PageRank personalizado, no domínio de Biomedicina. O trabalho utiliza dados para teste e uma fonte de conhecimento externa, que são reconhecidos por sua importância, tanto para o WSD como para o domínio em questão. Além dos procedimentos e resultados experimentais descritos no artigo, o software que implementa os experimentos e a fonte de conhecimento estão publicamente à disposição na Internet[4]. Por essas razões, uma pesquisa exploratória foi executada com o objetivo de reproduzir os experimentos e coletar resultados. Além disso, outro objetivo era identificar lacunas que pudessem ser exploradas. Nesse sentido, trabalhos relacionados que poderiam ser empregados de forma complementar à proposta de Agirre *et al.* foram investigados.

Dentre os trabalhos relacionados temos Navigli e Lapata [35, 36], abordados no Capítulo 3. Os autores apresentam um estudo sobre métricas de conectividade de grafos, para o WSD não supervisionado. A WordNet foi utilizada como fonte de conhecimento externo nesta pesquisa, e não estava voltada a um domínio específico. Dentre as métricas avaliadas, Degree e KPP obtiveram os melhores resultados. Os autores afirmam, em razão dos resultados, que a qualidade da fonte de conhecimento externa influencia diretamente a performance do WSD, afirmação esta que viremos a utilizar ao longo da tese. Além disso, as métricas experimentadas são independentes do léxico utilizado. O fato de induzirem um *ranking* de sentidos, empregando apenas a conectividade do grafo, torna possível a portabilidade entre algoritmos, línguas e fontes de conhecimento.

Considerando o trabalho de Agirre *et al.* com a abordagem baseada em grafos em um domínio específico, e o trabalho de Navigli e Lapata com a identificação da melhor abordagem baseada em grafos, em domínio independente, a seguinte hipótese foi levantada:

> **H1**: As métricas não-supervisionadas com melhor desempenho, encontradas em Navigli e Lapata [36] e Navigli e Lapata [35] levam ao melhor resultado no domínio da Biomedicina, em comparação ao cenário apresentado em Agirre *et al.* [3].

Para investigar a hipótese H1 foram estabelecidos objetivos. Eles incluem a elaboração de novos experimentos com a implementação de algoritmos. Para que os resultados da pesquisa possam ser comparados aos resultados de Agirre *et al.* [3], os objetivos contemplam os mesmos requisitos e meios de interpretação utilizados por esses autores. Sendo assim, temos os seguintes objetivos:

---

4 O software utilizado em [13] está disponível em http://ixa2.si.ehu.es/ukb/. O artigo referenciado dita o local onde os demais recursos podem ser encontrados. O UMLS e sua documentação estão à disposição em http://www.nlm.nih.gov/research/umls/.

1. reproduzir o experimento de Agirre *et al.* [3] empregando os mesmos recursos utilizados pelos autores ou, se não for possível, aqueles que se aproximem ao máximo das condições do experimento original;
2. implementar os algoritmos de KPP e Degree no software distribuído por Agirre *et al.*;
3. coletar e comparar os resultados obtidos.

A reprodução do experimento de Agirre *et al.*, que fez parte da pesquisa exploratória realizada no âmbito desta tese, utilizou um conjunto de instruções propostas pelos autores. Estas instruções e o material utilizado foram coletados do *site* http://ixa2.si.ehu.es/ukb/ em Julho de 2011. As etapas incluíam o *download* de código-fonte, arquivos de dados, softwares, configuração e compilação de ferramentas. A Figura 5.1 reúne todas essas etapas.

1. *download* e instalação do UMLS;
2. extração da tabela MRREL do UMLS;
3. *download* do *corpus* NLM-WSD;
4. *download* e instalação do MetaMap;
5. pré-processamento do *corpus* NLM-WSD, incluindo:
   - 5.1. anotação do *corpus* com o *parser* MetaMap,
   - 5.2. remoção de *stop phrases*,
   - 5.3. geração do dicionário,
   - 5.4. extração dos contextos;
6. compilação e execução.

Figura 5.1: Resumo das etapas do experimento de Agirre *et al.* [3]

Para o *download* são necessárias a solicitação e a aprovação de um registro junto ao *site* do *National Library of Medicine, Department of Health and Human Services* (NLM). Com o acesso, o *download* compreende a obtenção de cinco arquivos. Um deles (mmsys.zip) inclui um software para a navegação e extração das informações contidas no UMLS. Para extrair a tabela MMREL é necessária a configuração e utilização desse software. A tabela MMREL contém mais informações que as necessárias para o experimento. Por essa razão, um *script* extrai apenas as relações entre os conceitos, e as armazena em um arquivo no formato texto. O *corpus* corresponde a um conjunto de arquivos compactados em duas versões. Uma delas contém anotações no formato PMID (PubMed Identifier), que é utilizado nas demais etapas. O *parser* MetaMap está disponível em diferentes versões. Ele faz a anotação de textos de acordo com uma versão do UMLS. A versão do *parser* utilizada por Agirre *et al.* é a 2007, em razão de a versão do UMLS utilizada ser a 2007AB. No entanto, essa versão do *parser* não se encontra mais à disposição, o que estabeleceu um problema em potencial com as anotações do *corpus* e a tabela MMREL. Para contornar o problema, os autores do artigo foram contatados, e por intermédio deles se teve acesso à versão 2008 do *parser* MetaMap. Outra questão relacionada às versões dos softwares utilizados no experimento são os *scripts* elaborados pelos autores. Muitos deles são compatíveis apenas com as versões do período em que os experimentos foram executados. Em razão dessa limitação, não foi possível utilizar versões mais novas do MetaMap. Por outro lado, versões recentes do UMLS poderiam ser utilizadas, uma vez que as CUIs dos conceitos presentes no UMLS são únicas. Além disso, uma versão mais recente levaria (potencialmente) a resultados novos. Se a fonte de conhecimento é mais

recente mas o algoritmo (PageRank) é o mesmo, é possível avaliar brevemente as mudanças nos resultados ao longo das versões do grafo. Compreender esses resultados pode ou não confirmar a hipótese levantada por Navigli e Lapata [35], de que a fonte de conhecimento influencia os resultados. Portanto, foram empregadas no novo experimento as versões 2008 e 2011AA do MetaMap e do UMLS, respectivamente.

Todas as etapas foram executadas em consideração à janela de contexto padrão (20 conceitos, 10 antes e 10 após o conceito ambíguo). O software distribuído pelos autores, escrito em C++, foi então compilado. Este software executa duas etapas. A primeira utiliza o arquivo texto da tabela MMREL na geração de uma versão binária da mesma. O objetivo é reduzir o tempo de execução e otimizar o uso de memória. A segunda etapa utiliza a versão binária da MMREL, o dicionário de conceitos, os termos ambíguos e seus contextos para então desambiguá-los. Um esquema dessa etapa foi apresentado no *overview* da Figura 4.2. O resultado da reprodução do experimento levou a um percentual de acerto de 66,16%. Da mesma forma que Agirre *et al.* se posicionaram, os casos anotados como *none* não fazem parte desta análise de resultados.

5.1. Experimentos e Resultados

Com a reprodução do experimento concluída, o segundo objetivo envolve a implementação dos algoritmos KPP e *Degree*. A Figura 5.2 apresenta um *overview* completo dos experimentos. A codificação dos algoritmos segue as características descritas no Capítulo 3. Os experimentos realizados com os dois novos algoritmos utilizam os mesmos recursos e parâmetros empregados na reprodução do experimento com o PageRank. Agirre *et al.* [3] apresentam uma tabela contendo os resultados da desambiguação para cada conceito do *corpus*. Semelhante a esta, uma nova tabela foi elaborada (Tabela 5.1) com os resultados dos dois novos algoritmos. Os conceitos em itálico representam os casos difíceis discutidos no Capítulo 4. A coluna "#totalInst" representa a quantidade de instâncias avaliadas pelo algoritmo para cada conceito. Ou seja, ela representa todas as instâncias de conceitos ambíguos que não foram classificadas como *none* pelos anotadores. As colunas "#inst" representam a quantidade de instâncias classificadas corretamente por cada algoritmo. As colunas *percentual* (%) apresentam a taxa de instâncias classificadas corretamente, pela relação entre as colunas "#inst" e a coluna "#totalInst". Os melhores resultados estão indicados em negrito. Em alguns casos mais de um algoritmo atingiu o melhor resultado. A coluna "Agirre *et al.* (2010)" reproduz os resultados percentuais que o artigo apresenta para cada conceito. Não estão à disposição os valores absolutos de instâncias corretamente classificadas.

Como a Tabela 5.1 apresenta de forma discreta os resultados obtidos, outro meio de expor os resultados foi elaborado, através de um diagrama de Venn. A Figura 5.3 apresenta a interseção dos resultados de cada algoritmo em relação às instâncias. Dentre as 5.000 instâncias do *corpus*, 3.983 instâncias de conceitos foram avaliadas. O algoritmo PageRank personalizado classificou corretamente 2.635 instâncias. Deste total, apenas este algoritmo acertou o sentido correto de 580 instâncias (14,5%). O algoritmo KPP classificou corretamente 1.676 instâncias e destas, 676 (16,9%) foram acertadas apenas com este algoritmo. Por fim, 129 instâncias (3,2%) puderam ser classificadas apenas pelo algoritmo *Degree*, que por sua vez classificou 1.833 instâncias corretamente. Algumas instâncias foram classificadas corretamente por mais de um algoritmo. Nesta situação, 458 instâncias (11,4%) foram classificadas corretamente pelos algoritmos PageRank e KPP. Pelos algoritmos PageRank personalizado e *Degree* um total de 1.162 instâncias (29,1%) foram classificadas corretamente. KPP e *Degree* classificaram corretamente 107 instâncias (2,6%) do *corpus* NLM-WSD. Cerca de 435 instâncias (10,9%) foram classificadas corretamente por todos

os algoritmos. Por fim, nenhum dos três algoritmos conseguem classificar corretamente 436 instâncias (10,9%).

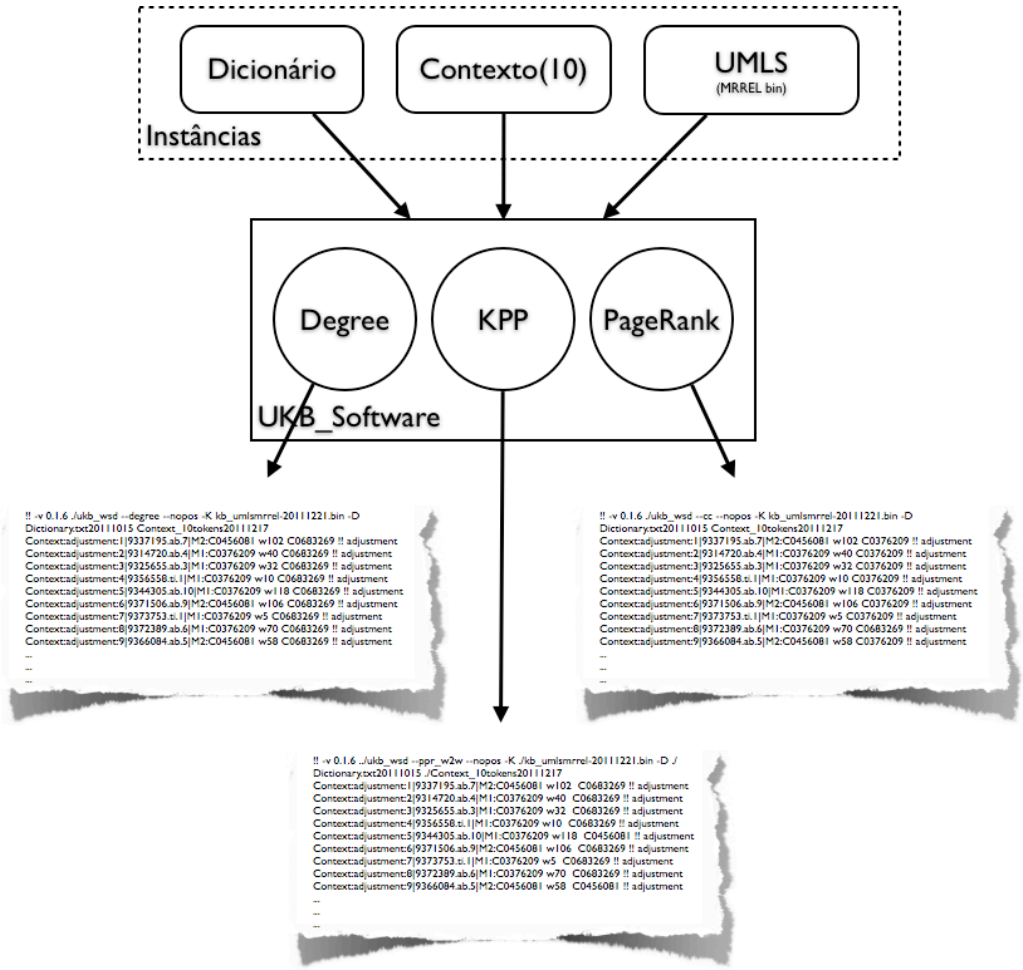

Figura 5.2: *Overview* do experimento com o modelo simples

Tabela 5.1: Resumo dos resultados

Conceito	#totalInst	Degree		KPP		PageRank		
		#inst	%	#inst	%	#inst	%	Agirre *et al.* [3]
adjustment	93	13	13,98	15	16,13	29	**31,18**	35,50
blood_pressure	100	18	18,00	44	44,00	47	**47,00**	48,00
cold	95	12	12,63	5	5,26	24	**25,26**	28,40
condition	92	3	3,26	90	**97,83**	33	35,87	48,90
culture	100	11	11,00	16	**16,00**	14	14,00	77,00
degree	65	2	3,08	2	3,08	61	**93,85**	93,80
depression	85	85	**100,00**	4	4,71	61	71,76	94,10
determination	79	0	0,00	36	**45,57**	36	**45,57**	94,90
discharge	75	14	18,67	1	1,33	53	**70,67**	69,30
energy	100	97	97,00	42	42,00	98	**98,00**	27,60
evaluation	100	55	**55,00**	51	51,00	52	52,00	50,00
extraction	87	16	18,39	82	**94,25**	22	25,29	27,60
failure	29	4	13,79	6	20,69	8	**27,59**	72,40
fat	73	71	**97,26**	3	4,11	70	95,89	95,90
fit	18	0	0,00	18	**100,00**	2	11,11	11,10
fluid	100	100	**100,00**	0	0,00	90	90,00	92,00
frequency	94	34	36,17	94	**100,00**	92	97,87	98,90
ganglion	100	23	23,00	71	71,00	77	**77,00**	64,00
glucose	100	86	86,00	10	10,00	90	**90,00**	90,00
growth	100	37	37,00	63	**63,00**	37	37,00	37,00
immunosuppression	100	59	**59,00**	44	44,00	57	57,00	62,00
implantation	98	37	37,76	22	22,45	86	**87,76**	84,70
inhibition	99	1	1,01	95	**95,96**	14	14,14	22,20
japanese	79	18	22,78	6	7,59	72	**91,14**	64,60
lead	29	27	**93,10**	27	**93,10**	27	**93,10**	93,10
man	92	24	26,09	14	15,22	32	**34,78**	44,60
mole	84	0	0,00	83	**98,81**	21	25,00	27,40
mosaic	97	52	53,61	24	24,74	77	**79,38**	66,00
nutrition	89	20	22,47	22	24,72	29	**32,58**	32,60
pathology	99	83	**83,84**	78	78,79	41	41,41	28,30
pressure	96	96	**100,00**	0	0,00	91	94,79	97,90
radiation	98	61	**62,24**	61	**62,24**	61	**62,24**	53,10
reduction	11	2	18,18	2	18,18	5	**45,45**	54,50
repair	68	43	63,24	48	70,59	53	**77,94**	76,50
resistance	3	3	**100,00**	3	**100,00**	3	**100,00**	66,70
scale	65	0	0,00	40	61,54	50	**76,92**	84,60
secretion	100	97	97,00	1	1,00	99	**99,00**	99,00
sensitivity	51	2	3,92	46	**90,20**	24	47,06	27,50
sex	100	32	32,00	35	35,00	88	**88,00**	85,00
single	100	9	9,00	6	6,00	89	**89,00**	82,00
strains	93	4	4,30	92	**98,92**	63	67,74	96,80
support	10	8	**80,00**	2	20,00	8	**80,00**	80,00
surgery	100	89	89,00	21	21,00	98	**98,00**	97,00
transient	100	99	**99,00**	1	1,00	98	98,00	99,00
transport	94	93	**98,94**	93	**98,94**	93	**98,94**	69,10
ultrasound	100	84	**84,00**	16	16,00	84	**84,00**	83,00
variation	100	20	20,00	79	79,00	80	**80,00**	75,00
weight	53	25	47,17	24	45,28	35	**66,04**	56,60
white	90	64	**71,11**	38	42,22	61	67,78	63,30
Soma #inst	**3983**	**1833**		**1676**		**2635**		
Média			**46,02**		**42,08**		**66,16**	**65,89**

5.2. Discussão

Em primeiro lugar vamos considerar algumas particularidades a respeito dos valores relacionados na Tabela 5.1.

O resultado obtido com a reprodução do experimento de Agirre *et al.* levou a um melhor resultado (66,16%) do que aquele atingido no experimento conduzido pelos autores (65,89%). Não foi possível encontrar explicações conclusivas, mas tudo indica que dois fatores podem ser responsáveis por essa diferença. O primeiro é que os parâmetros utilizados na ferramenta de extração da tabela MRREL podem não ser os mesmos. Não há documentação precisa a respeito de quais vocabulários deveriam ser selecionados. O segundo fator é que o UMLS sofre pequenas atualizações entre a versão utilizada pelos autores e a que foi utilizada na reprodução. Desta forma a estrutura do grafo e, consequentemente, os relacionamentos entre conceitos, foram alterados. Os demais algoritmos não obtiveram um resultado geral melhor do que aquele alcançado pelo PageRank.

Alguns conceitos têm menos de 20% de suas instâncias avaliadas pelos algoritmos. Curiosamente, algumas delas não fazem parte do conjunto de conceitos difíceis, como é o caso de *fit*, *reduction* e *resistance*. Isso significa que, apesar de não serem consideradas difíceis em Weeber *et al.* [53], grande parte das instâncias anotadas não obtiveram classificação. A taxa média geral de utilização das instâncias é de 81,29%, enquanto a dos conceitos considerados difíceis é de 78,33%. Essa diferença indica que a existência de uma grande quantidade de instâncias anotadas pelos anotadores como *none* não significa que as mesmas são consideradas difíceis. Além disso, a classificação dos conceitos considerados difíceis obteve resultados diferentes do âmbito global. Enquanto o algoritmo de KPP obteve uma taxa de 48,83% (+6,75 pontos que no geral), os algoritmos PageRank e Degree têm a performance reduzida a 51,06% (-15,1 pontos) e 27,02% (-19 pontos), respectivamente. Uma explicação para o fato de que KPP tenha um melhor desempenho é a relação entre a dificuldade dos anotadores em escolher um conceito, e o nível de centralidade do conceito correto no contexto avaliado. O conceito mais central, na janela de contexto em que se encontra, leva à classificação correta de um conceito ambíguo difícil.

Outra questão relacionada aos resultados da Tabela 5.1 são as variações de resultados entre os algoritmos. Dentre os melhores resultados, em treze conceitos (26% do total) um único algoritmo obteve o resultado maior ou igual ao dobro dos outros algoritmos. Por exemplo, para o conceito *fit* o algoritmo KPP classificou corretamente 100% das instâncias analisadas, e o PageRank apenas 11,1%. Ao contrário do discutido em Agirre *et al.* (2010), o fator determinante para as escolhas na classificação não é a densidade com a qual o sentidos estão conectados. O algoritmo de KPP destaca aqueles que são centrais na estrutura do grafo, e não apenas pela densidade dos relacionamentos. Este comportamento levou a um efeito contrário com KPP, onde o conceito *secretion* chegou ao pior resultado dos três algoritmos (1% de acerto). Em resumo, o algoritmo PageRank obteve 62% (8) destes melhores resultados, enquanto KPP chegou a 38% (5). O algoritmo Degree não se destacou em nenhum dos conceitos.

A conclusão alcançada com a análise dos resultados dos experimentos desenvolvidos neste trabalho não confirma a hipótese H1. Os algoritmos discutidos em Navigli e Lapata [36] e Navigli e Lapata [35], cujo desempenho se destacou em experimentos sem domínio específico, não repetiram a mesma performance no domínio específico da Biomedicina.

Todos esses aspectos ligados às variações de resultados entre algoritmos e conceitos levaram a dúvidas em relação à performance em nível das instâncias. Se alguns dos algoritmos podem ter resultados muito ruins ou muito bons em relação aos demais, torna-se necessário identificar a proporção e a distribuição desses resultados. A Figura 5.3 apresenta a distribuição das 3.983

instâncias classificadas nos experimentos. Dentre aquelas que foram corretamente classificadas, o resultado de cada algoritmo vs. instância permite estabelecer um conjunto de considerações.

Em primeiro lugar, apesar de o algoritmo de PageRank obter o melhor resultado geral (Tabela 5.1), o algoritmo KPP classificou exclusivamente o maior número de instâncias. Foram 676 casos (16,94% das 3983 instâncias) contra 580 do algoritmo PageRank (14,56%). Por outro lado, o algoritmo *Degree* classificou corretamente cerca de 60% das instâncias (1597) classificadas pelo PageRank. Conforme Navigli e Lapata [35], a complexidade dos algoritmos PageRank e Degree é, respectivamente $O(n^2)$ e $O(n)$. Isso significa que mais da metade das instâncias pode ser analisada em um período de tempo menor, se for utilizado o algoritmo Degree, acarretando num melhor desempenho. Dentre os 435 casos em que todos os algoritmos identificaram corretamente o sentido, apenas 3 conceitos (em itálico na Figura 5.4) fazem parte do conjunto considerado difícil. Os conceitos *lead*, *resistance* e *transport* tiveram aproximadamente 100% das instâncias classificadas corretamente pelos três algoritmos.

A união dos resultados corretos dos três algoritmos (PageRank ∪ KPP ∪ *Degree*) corresponde a um total de 3.547 instâncias classificadas corretamente. Essa quantidade corresponde a uma taxa de acerto de 89,05%. Esse resultado leva a crer que, ao invés de se tentar corrigir os erros na classificação, é possível superar os resultados encontrados até o momento, com o emprego de múltiplos métodos. Tal hipótese foi levantada, mas não confirmada, nas conclusões de Navigli e Lapata [36]. Os autores colocam que a performance poderia crescer se houvesse um *framework* que escolhesse o algoritmo mais adequado.

adjustment (1/93)	fat (1/73)	radiation (61/98)
blood pressure (1/100)	frequency (34/94)	repair (39/68)
culture (7/100)	ganglion (8/100)	**resistance** (3/3)
depression (3/85)	glucose (1/100)	sex (4/100)
discharge (1/75)	*immunosuppression* (6/100)	strains (3/93)
energy (39/100)	implantation (6/98)	surgery (19/100)
evaluation (2/100)	japanese (1/79)	**transport** (93/94)
extraction (3/87)	**lead** (27/29)	weight (17/53)
failure (4/29)	pathology (24/99)	white (26/90)

Figura 5.4: Lista de conceitos classificados corretamente
por todos os algoritmos (#inst / #totalInst)

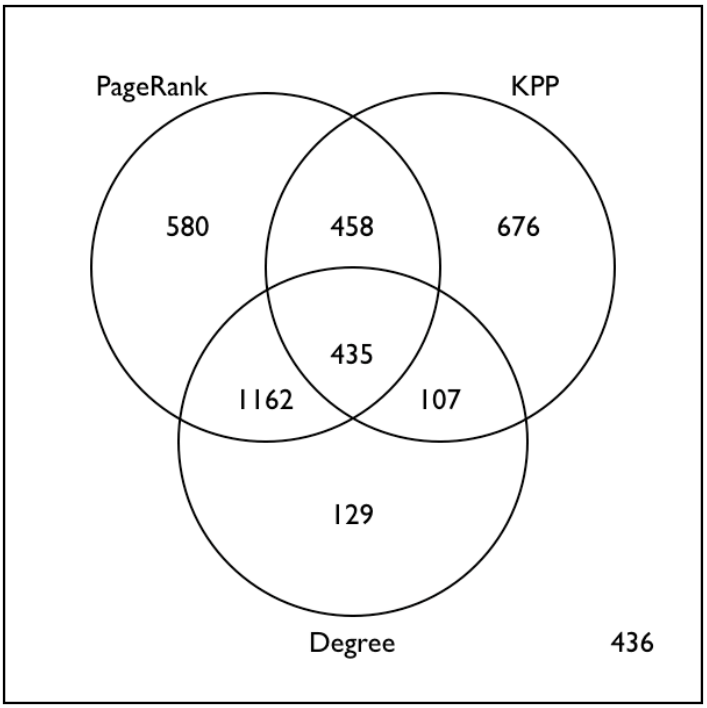

Figura 5.3: Acertos na distribuição das instâncias por métricas

6. MODELO HÍBRIDO DE MÉTRICAS

Inspirado na sugestão inicial de Navigli e Lapata [35, 36], acrescida dos resultados de Agirre *et al.* [3], e fundamentado nos experimentos prévios realizados, este trabalho propõe um modelo híbrido com o emprego de métricas na seleção do sentido para um dado conjunto de instâncias ambíguas. Os resultados discutidos até aqui (Seção 5.2) demonstram que, se a métrica certa for selecionada, o desempenho pode aumentar, seja ele em termos de percentual de acerto ou em termos de ganhos de processamento.

Considerando um processo de cinco etapas para desambiguar instâncias (Figura 6.1), o modelo híbrido necessita de uma ou mais *features* e heurísticas para selecionar uma métrica. Nesse processo a primeira etapa (Figura 6.1.a) diz respeito à seleção das instâncias. Somente aquelas consideradas relevantes são utilizadas. Por exemplo, nos experimentos com NLM-WSD as instâncias classificadas como *none* são desconsideradas. A segunda etapa (Figura 6.1.b) compreende a extração de *features*. As *features* correspondem a informações a respeito das instâncias, que serão utilizadas na etapa seguinte (Figura 6.1.c). Detalhe, esta etapa não deve substituir a tarefa da métrica de identificar o sentido correto (apontada na Figura 6.1.d) mas, sim, servir à seleção da métrica mais adequada para essa análise.

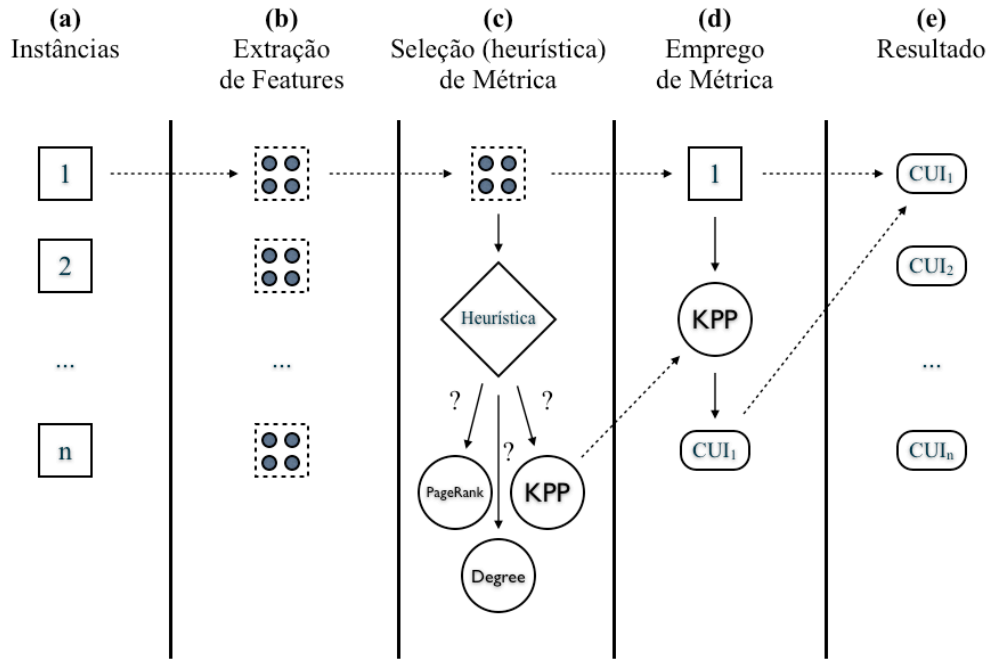

Figura 6.1: Etapas do estudo experimental

A etapa de seleção de métrica (Figura 6.1.c) compreende a escolha de algum método que, utilizando as *features* selecionadas na etapa anterior, selecione a métrica que irá classificar uma determinada instância. Com a métrica selecionada, o processo de classificação das instâncias é o mesmo do modelo simples. O sistema que implementa o modelo simples de WSD pode ser complementado, então, com este processo de extração e classificação de instâncias (Figura 6.2).

O modelo híbrido de métricas é então avaliado de duas formas. A primeira é a comparação dos resultados obtidos por conceito do NLM-WSD em relação aos obtidos nos experimentos deste trabalho e aqueles obtidos nos demais trabalhos apresentados anteriormente. Ou seja, os resultados

são analisados em comparação àqueles descritos na Tabela 5.1. A segunda forma de avaliação compreende a análise dos resultados em nível das instâncias. De forma semelhante à análise feita na Seção 5.1, a distribuição dos erros e acertos encontrados neste estudo experimental é empregada na avaliação dos resultados obtidos com diferentes propostas de configuração de *features* e métodos na seleção de métricas. Essas configurações exigiram a elaboração de experimentos que combinassem as opções de *features* e heurísticas de seleção de métricas. Tais experimentos foram conduzidos no mesmo procedimento e rigor metodológico que os anteriores, e são registrados e analisados no presente trabalho. Ao final deste processo, foi estabelecido um modelo para o emprego de múltiplas métricas baseadas em grafos para o WSD.

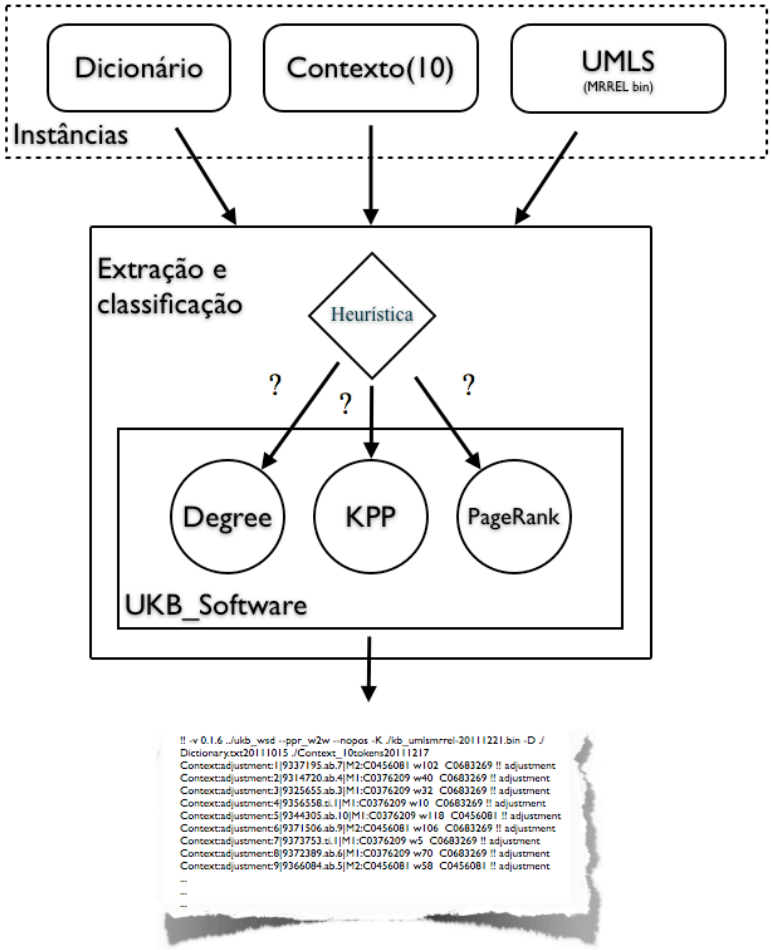

Figura 6.2: *Overview* do modelo híbrido resultante

6.1. Seleção de *Features*

Esta proposta estabelece o uso de um conjunto de *features*. As *features* podem ter origem com base nos dados e resultados utilizados nos experimentos, e serem selecionadas a partir de análise estatística. Por outro lado, as *features* podem ser obtidas do *corpus* ou das fontes de conhecimento externas, não utilizadas nos experimentos.

Considerando cada instância do *corpus*, temos à disposição:

- palavra e CUI para os candidatos a sentido;
- palavra e CUI para o conceito correto;
- fonte do conceito correto e dos candidatos;

- tipos semânticos do conceito correto e dos candidatos;
- categoria morfossintática do conceito correto e dos candidatos;
- conjunto de métricas que identificam o sentido de uma instância;
- quantidade de palavras (título, sentença e resumo) de uma instância ambígua.

Essas informações foram reunidas e analisadas estatisticamente para a elaboração de heurísticas (tema discutido na Seção 6.2) ou para servir como entrada a um algoritmo de aprendizado de máquina. Observa-se que não foram selecionadas *features* que são empregadas diretamente pelas métricas. Em outras palavras, foram selecionadas *features* que não fazem parte do processo direto de análise das instâncias. O objetivo é evitar o processamento redundante de informações e não utilizar informações que estabeleçam algum tipo de tendência para uma métrica ou candidato na etapa seguinte. Um exemplo é a utilização do grau dos candidatos no grafo do contexto das instâncias, utilizado pela métrica Degree. Fazer uso dessa informação pode beneficiar a indicação das instâncias em que Degree identifica corretamente o sentido mas não permite que os demais, reconhecidos pelas outras métricas, possam ser identificados. Apesar de ser uma estratégia possível de ser trabalhada, a combinação de heurísticas dessa natureza não foi adotada neste trabalho.

A seleção de *features* do ponto de vista estatístico utiliza a frequência dos conceitos corretos, métricas e demais *features*. Além disso, a coocorrência entre elas também foi calculada. As estatísticas levantadas a partir do *corpus*, do UMLS e dos resultados dos experimentos permitem estabelecer as bases da criação de heurísticas de seleção de métricas. Por exemplo, a correlação entre uma métrica e um conceito pode estabelecer uma regra para seleção da métrica a partir das características desse conceito. Tais características podem ser encontradas nas fontes de conhecimento, sejam externas ou não. Essas ideias são exploradas na Seção 6.2.

A consulta de mais informações a respeito dos conceitos viabiliza uma perspectiva generalizada sobre os mesmos. Por exemplo, para cada instância ambígua do NLM-WSD há um conjunto de conceitos candidatos presente no UMLS. Esses candidatos foram estabelecidos na construção do *corpus* e estão relacionados a CUIs do metatesauro. Cada CUI, por sua vez, está associada à fonte original da qual os conceitos foram obtidos na construção do metatesauro (tabela MRCONSO.RRF). A fonte pode ser identificada para todos os candidatos associados a uma instância, ou apenas para o candidato correto. A fonte dos candidatos pode então ser empregada como uma *feature* para agregar valores estatísticos. Por exemplo, ao invés de contabilizar a coocorrência entre métrica e candidato correto, são utilizadas a métrica e a fonte do candidato correto. Dadas estas constatações, a fonte original de cada conceito se torna uma alternativa de *feature*. Este método é empregado no modelo híbrido (formalizado na Seção 6.3).

Um experimento foi elaborado com o objetivo de identificar as *features* mais relevantes para a escolha da métrica. Dois algoritmos de aprendizado de máquina com diferentes parâmetros foram utilizados (J48 e *Naïve Bayes Tree*)[5]. Como candidatos a *features* foram selecionados, para cada instância, a fonte do conceito correto e até três tipos semânticos do conceito correto, e a contagem de palavras da sentença, do título e do resumo em que o conceito ambíguo ocorre. O atributo classe selecionado é o das métricas que identificaram o sentido correto no experimento preliminar. O melhor resultado do experimento (Tabela 6.1 e Figura 6.3) revelou por meio de uma árvore de decisão[6] (J48 com poda 32) que a fonte do conceito correto se correlaciona com a métrica de melhor desempenho. Essa constatação determinou a escolha do uso das fontes dos candidatos como um meio de selecionar a melhor métrica para o WSD.

---

5 Os experimentos foram implementados com a ferramenta Weka (www.cs.waikato.ac.nz/ml/weka/) na sua versão 3.6.7.

6 Esse experimento também fez parte do processo de análise de alternativas para seleção de métricas (descrito na Seção 6.2), por esta razão constam também os algoritmos NBTree e SMO(SVM). A árvore extraída, além de informações complementares, pode ser encontrada em http://www.rodrigo.goulart.nom.br/tese/

Por fim, dentre as categorias morfossintáticas da instância e seu contexto, apenas a informação de que os conceitos são substantivos consta no NLM-WSD. O texto poderia ser processado por outras ferramentas de anotação e assim determinar-se a estrutura sintática das sentenças em que as instâncias estão inseridas. No entanto, a estrutura sintática das sentenças não representa necessariamente a mesma estrutura de conceitos analisada pelas métricas. Portanto, optou-se pela não utilização deste novo processo nem dessas informações.

Tabela 6.1: Testes com aprendizado de máquina

Algoritmo		**%acerto**
J48	**Poda**	
	2	69,24
	4	70,54
	8	71,34
	16	72,44
	32	**72,6**
	64	71,24
	128	67,9
	256	46,96
	512	47,12
Naïve Bayes		66,76
SMO(SVM)		70,96
NBTree		71,78

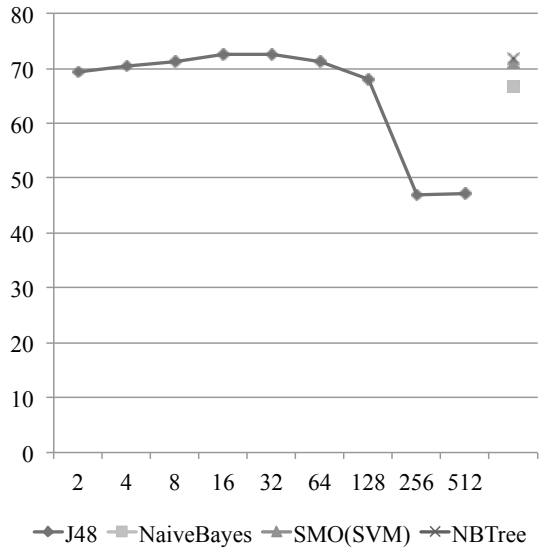

Figura 6.3: Testes com os demais algoritmos

Concluindo, o trabalho passa a estudar a seleção de métricas com o emprego das fontes dos candidatos como feature, além das métricas que identificam o sentido de cada instância. Essas informações são utilizadas na etapa seguinte, a seleção de métricas, abordada na próxima seção. As fontes de candidatos que fazem parte desta pesquisa são apresentadas na Seção 6.3.

6.2. Seleção de Métricas

Para a seleção de métricas, duas alternativas foram exploradas: o aprendizado de máquina e a probabilidade da correlação entre métricas e *features*.

A utilização de aprendizado de máquina busca estabelecer um modelo de aprendizado capaz de classificar as instâncias ambíguas com métricas. Neste caso, considerando as *features* propostas na Seção 6.2, quatro algoritmos para o aprendizado de máquina foram utilizados (J48, Naïve Bayes, Naïve Bayes Tree e SVM). Os conjuntos de treino e teste foram elaborados com base nos resultados da classificação obtidos nos experimentos da Seção 5.1, e que correspondem às etapas (d) e (e) da Figura 6.1. Desta forma, as *features* de cada instância do *corpus* NLM-WSD são consideradas atributos de um vetor, semelhante ao apresentado na Seção 2.3, Figura 2.4. A métrica ou as métricas que identificam o sentido correto da instância NLM-WSD são o atributo classe da instância para o aprendizado de máquina. Assim pode ser estabelecido um conjunto de treino e teste que classifique qual a melhor ou melhores métricas para uma dada instância. Por exemplo, considere o conjunto de vetores (*features* e atributo classe) na Figura 6.4. Cada vetor representa uma instância do NLM-WSD que foi classificada corretamente por uma ou mais métricas.

Experimentos com aprendizado de máquina foram elaborados para seleção de métricas. Utilizando as *features* propostas na Seção 6.1, duas configurações foram estabelecidas. Elas compreendem subconjuntos de *features* e níveis de poda no algoritmo J48. Uma delas é o experimento utilizado na seleção de *features* (também apresentado anteriormente na Seção 6.1). A outra são experimentos que empregam o conjunto das fontes dos candidatos como *features*. Ao invés de tratar cada fonte de candidato como uma *feature* (individualmente), o conjunto das *features* foi retratado como uma única *feature*. Por exemplo, considere que para uma dada instância *i* as fontes dos seus candidatos a sentido são AOD e MTH (estas siglas e as respectivas fontes são apresentadas na Seção 6.3). Ao invés de considerarmos cada fonte separadamente como uma *feature*, o conjunto das fontes AOD-MTH foi utilizado como uma única fonte. O objetivo dessa modificação é identificar se os conjuntos de fontes podem conduzir à escolha da métrica com generalização dos casos em uma única *feature*. Todo esse processo será detalhado na Seção 6.3.

Instância	*feature* 1	*feature* 2	…	*feature* n	classe
1	-	X	X	X	PageRank
2	-	-	X	X	PageRank
3	X	-	-	-	KPP
4	X	X		X	KPP+Degree
5		X	X	X	KPP+Degree +PageRank
6	X	-	X	X	-
…	…	…	…	…	...

Figura 6.4: Exemplo de vetor de *features*

Os resultados dos experimentos com aprendizado de máquina revelaram dois problemas. O primeiro é que o modelo de aprendizado extraído com melhor resultado/configuração utiliza a fonte do conceito correto, além de eventualmente outras *features*, para determinar a(s) métrica(s) mais adequada(s). Ele corresponde ao modelo criado para seleção de *features*. O problema é que identificar o conceito correto é a etapa seguinte no processo de desambiguação (descrito na Figura 6.1) e portanto não poderia ser utilizado na etapa de seleção de métrica. A segunda situação é a de que configurações que não têm problemas desse tipo tiveram resultados inferiores ao da melhor métrica e por essa razão foram descartados. Os resultados ficaram abaixo dos até aqui levantados (cerca de 30% de taxa de acerto).

A probabilidade da correlação é outro meio de selecionar uma métrica para o WSD. Nesse caso, a correlação entre o valor de uma *feature* e as métricas disponíveis pode determinar qual a métrica mais adequada para a classificação. Este estudo considerou o emprego da fonte dos candidatos como a melhor alternativa. Os experimentos da Seção 6.1 identificaram a fonte como a *feature* mais promissora e por essa razão foi utilizada. Considerando a fonte do candidato correto e a(s) métrica(s) que o identificam é possível estabelecer uma correlação entre a estrutura da fonte do candidato e esta(s) métrica(s). Utilizando o conjunto de todas as instâncias disponíveis no *corpus* NLM-WSD é possível então estabelecer a probabilidade da correlação entre fonte e métrica. Em outras palavras, ela estabelece um relação entre fonte e métrica baseada na probabilidade identificada a partir dos exemplos do *corpus*. Portanto, para cada fonte de candidato é possível

sugerir a(s) métrica(s) com maior probabilidade de identificar o conceito correto. No entanto, essa sugestão se baseia apenas num candidato. Considerando o fato de que as instâncias de um *corpus* podem ter um ou mais candidatos a sentido, de diferentes fontes, se for determinada uma métrica para cada candidato é necessária alguma heurística para selecionar a métrica definitiva. O método mais simples é utilizar a métrica do candidato com a probabilidade mais alta. Os experimentos indicaram que a taxa de acerto desse método é a mesma da melhor métrica no modelo simples (66,03 % de acerto).

Para solucionar o problema, o conjunto das fontes de candidatos foi empregado para seleção da métrica. Considerar as fontes como apenas uma *feature* viabilizou a construção do modelo e obtenção de melhores resultados em comparação ao modelo simples. A Seção 6.3 descreve detalhes de como a seleção e extração das features foi implementada. Descreve ainda como a heurística de seleção é estabelecida e os resultados são atingidos, com o modelo híbrido.

6.3. O Modelo

Este trabalho propõe um modelo híbrido para a seleção de métricas para o WSD baseado em grafos. Para uma dada instância ambígua, a seleção da métrica que vai ser aplicada é determinada por sua probabilidade condicional. A probabilidade de uma métrica nesse caso é dependente da ocorrência de um conjunto de fontes de candidatos a sentido para a instância analisada.

As instâncias do *corpus* NLM-WSD anotadas com o conceito correto no UMLS também foram associadas a um conjunto de conceitos candidatos a sentido correto. Todos estes conceitos possuem CUIs, que por sua vez estão associadas às suas fontes (vocabulário de origem). Utilizando os resultados do experimento preliminar, resumido na Tabela 5.1, é estabelecida uma medida de probabilidade entre as métricas que identificam corretamente o sentido de uma instância. O modelo proposto foi avaliado em comparação com os resultados do experimento preliminar, e os resultados dessa avaliação são apresentados a seguir.

Para descrever o processo de seleção de métrica a partir das fontes, considere:

- F = o conjunto de todas as fontes dos conceitos presentes no UMLS.
- I = o conjunto das instâncias de palavras ambíguas do *corpus* NLM-WSD;
- i = uma instância de conceito ambíguo, onde $i \in I$;
- S_i = o conjunto dos candidatos a sentido da instância i;
- f_i = o conjunto da união das fontes dos candidatos da instância i;
- F_t = o conjunto da união de todos os conjuntos de candidatos das instâncias pertencentes a I;
- M = o conjunto das métricas, neste caso { Deg, Kpp, Ppr};
- M_i = conjunto das métricas que identificam o sentido da instância i.

O conjunto das fontes F inclui todas as fontes utilizadas nos experimentos (relembrando, as fontes são provenientes de bases bem constituídas que descrevemos logo a seguir). Portanto, temos :

F = { AOD, CHV, MSH, MTH, NCI, NDFRT, SCTSPA, SNOMEDCT }.

AOD (*Alcohol and Other Drug Thesaurus*)[7] é um guia de conceitos para pesquisadores e profissionais na área de álcool e outras drogas, mantido pelo *National Institute on Alcohol Abuse and Alcoholism* (NIAAA). Serve como um vocabulário controlado para indexação e recuperação de informação em sistemas de banco de dados. CHV (*Consumer Health Vocabulary*)[8] é produzido pelo *Biomedical Informatics Department* da Universidade de Utah em colaboração com outras quatro instituições. Seu objetivo é permitir a transcrição automática de conceitos técnicos e termos simples para leigos. MSH (*Medical Subject Headings*)[9] é um tesauro para indexação, catalogação e busca de informações e documentos sobre Biomedicina e saúde. Ele é mantido pela *National Library of Medicine* e é utilizado, por exemplo, no sistema de busca do site PubMed para pesquisas por assuntos. MTH (*Unified Medical Language System® Metathesaurus*)[10] contém conceitos, relacionamentos e outras informações utilizadas pela *National Library of Medicine* para facilitar a construção do UMLS. NCI (*National Cancer Institute*)[11] reúne conceitos relacionados ao câncer no atendimento clínico, pesquisa e atividades administrativas. NDFRT (*National Drug File - Reference Terminology*)[12] mantido pelo U.S. *Department of Veterans Affairs, Veterans Health Administration*. Ele é utilizado na classificação de drogas em termos de seus ingredientes, estrutura química, entre outras. SNOMEDCT (*Systematized Nomenclature of Medicine-Clinical Terms*)[13] é mantido pelo *Standards Development Organisation*. Ele reúne conceitos para a padronização de registros médicos empregados internacionalmente.

O conjunto *I* inclui todas as instâncias cujo sentido foi identificado pelos anotadores do NLM-WSD. Da mesma forma que no experimento preliminar, as instâncias classificadas pelos anotadores como *none* não foram consideradas. O conjunto final leva a um total de 3983 instâncias. Contudo, apenas as instâncias cujo sentido foi identificado por pelo menos uma métrica foram consideradas. Como se deseja analisar apenas a probabilidade das métricas, as instâncias cujo sentido não foi identificado por nenhuma métrica foram descartadas. Portanto, restaram 3547 instâncias em *I*. Cada instância *i* tem associado a ela um conjunto de conceitos candidatos a sentido S_i. Como estes candidatos não consideram o contexto em que a instância se encontra, o conjunto dos candidatos de um conceito é igual para todas as suas instâncias. Por exemplo, o conjunto dos candidatos a sentido do conceito *cold* é S_{cold} = { C0024117, C0009264, C0234192, C0009443, C0010412 } e as suas fontes são respectivamente { SNOMEDCT, SNOMEDCT, SCTSPA, SNOMEDCT, SNOMEDCT }. Assim, F_{cold} = { SCTSPA, SNOMEDCT }. Temos então o conjunto dos conjuntos de fontes de candidatos:

7 http://etoh.niaaa.nih.gov/AODVol1/aodthome.htm (Último acesso em 6 de Fevereiro de 2013).

8 http://consumerhealthvocab.org/ (Último acesso em 6 de Fevereiro de 2013).

9 http://www.nlm.nih.gov/mesh/meshhome.html (Último acesso em 6 de Fevereiro de 2013).

10 http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MTH/ (Último acesso em 6 de Fevereiro de 2013).

11 http://ncit.nci.nih.gov/ (Último acesso em 6 de Fevereiro de 2013).

12 http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/ (Último acesso em 14 de Maio de 2013).

13 http://www.ihtsdo.org/snomed-ct/ (Último acesso em 14 de Maio de 2013).

F_t = { { AOD, MTH }, { CHV }, { CHV, MTH }, { CHV, SCTSPA }, { CHV, SNOMEDCT }, { MSH }, { MSH, SCTSPA }, { MSH, SCTSPA, SNOMEDCT }, { MSH, SNOMEDCT }, { MTH, SCTSPA }, { MTH, SNOMEDCT }, { NCI }, { NCI, SCTSPA }, { NCI, SNOMEDCT }, { NDFRT, SCTSPA }, { SCTSPA }, { SCTSPA, SNOMEDCT }, { SNOMEDCT } }

Cada conceito está associado a um conjunto de fontes de F_t. A Tabela A.1 do Apêndice A apresenta a relação entre conceitos e os conjuntos de fontes.

A partir dessas definições e análises a probabilidade condicional de uma métrica *m*, dado um conjunto de fontes $f \in F_t$, é determinada por:

$$P(m \mid f) = \frac{P(m \cap f)}{P(f)}, \text{ onde}$$

$$P(m \cap f) = \frac{|m \cap f|}{|I|} \text{ e } P(f) = \frac{|I_f|}{|I|}$$

Para cada instância *i* do espaço amostral *I* há um conjunto de métricas M_i que identifica o sentido correto de *i.* A probabilidade condicional $P(m \mid f)$ é dada pela divisão de dois valores. O primeiro é a probabilidade de a métrica *m* ocorrer com o conjunto de fontes *f*, dada por $P(m \cap f)$. Esta probabilidade é calculada pelo número de instâncias de *I* em que a interseção ocorre, dividido pelo número de instâncias de *I*. O outro valor é a probabilidade da ocorrência do conjunto de fontes *f*, dado por $P(f)$. Ele é determinado pelo número de instâncias de *i* em que *f* ocorre.

Por exemplo, considere a métrica Degree e o conjunto de fontes F_i = { SCTSPA, SNOMEDCT }. Em cerca de 218 instâncias do *corpus* NLM-WSD (instâncias do espaço amostral *I*) o sentido foi identificado pela métrica Degree e utilizava candidatos cujas fontes correspondem a *f*. Considerando o total de 3547 instâncias, a probabilidade de $P($ Degree $\cap$ { *SCTSPA* , *SNOMEDCT* }) é 0,061460389. A probabilidade deste conjunto de fonte é determinada pela frequência com que foi empregado. Cerca de 521 instâncias em *I* têm como fontes de candidatos o conjunto { SCTSPA, SNOMEDCT }. Uma lista completa das ocorrências, da interseção entre métricas e conjuntos de fontes, e das probabilidades das fontes pode ser encontrada no Apêndice B, tabelas B.1, B.2 e B.3, respectivamente.

Portanto, $P($ { SCTSPA, SNOMEDCT }) = 521 / 3547 = 0,146884691. Este, inclusive, é o conjunto mais frequente de fontes. Considerando *M*, o conjunto das métricas, e *F*, o conjunto dos conjuntos das fontes, é possível determinar a probabilidade condicional para todas as métricas em relação a todos os conjuntos de fontes. Com esta relação a métrica com maior probabilidade pode ser selecionada para analisar os candidatos de cada instância de *I* em seu contexto. A Tabela 6.2 resume esses resultados.

Cada métrica pode então ser analisada a partir da sua probabilidade condicional em razão de um conjunto de fontes pré-estabelecido. A métrica cuja probabilidade é a mais alta (identificada em negrito na Tabela 6.2) é então selecionada como a mais indicada para classificação de uma instância com o dado conjunto de fontes. Por exemplo, considere uma instância *i* cuja palavra ambígua é *adjustment*. Os conceitos candidatos são *individual adjustment*, *psychological adjustment* e *adjustment action*. As fontes desses conceitos são respectivamente SNOMEDCT, MSH SCTSPA.

Considerando o conjunto das fontes { MSH, SCTSPA, SNOMEDCT } a métrica recomendada é o Ppr (PageRank Personalizado).

Tabela 6.2: Resumo das probabilidades condicionais entre métricas e conjuntos de fontes

Conjunto das Fontes	Deg	Kpp	Ppr	Métrica selecionada
{ AOD, MTH }	**1**	**1**	**1**	Deg
{ CHV }	0,4	0,6	**0,8**	Ppr
{ CHV, MTH }	0,37	**0,63**	0,37	Kpp
{ CHV, SCTSPA }	0,457627119	0,440677966	**0,842615012**	Ppr
{ CHV, SNOMEDCT }	0,48241206	0,457286432	**0,75879397**	Ppr
{ MSH }	0,898989899	0,212121212	**0,98989899**	Ppr
{ MSH, SCTSPA }	0,450381679	0,557251908	**0,636132316**	Ppr
{ MSH, SCTSPA, SNOMEDCT }	0,295454545	0,340909091	**0,659090909**	Ppr
{ MSH, SNOMEDCT }	0,674157303	0,171348315	**0,834269663**	Ppr
{ MTH, SCTSPA }	0,391304348	**0,601449275**	0,550724638	Kpp
{ MTH, SNOMEDCT }	0,367346939	0,476190476	**0,68707483**	Ppr
{ NCI }	**1**	**1**	**1**	Deg
{ NCI, SCTSPA }	0,234693878	0,724489796	**0,785714286**	Ppr
{ NCI, SNOMEDCT }	0,248275862	**0,848275862**	0,482758621	Kpp
{ NDFRT, SCTSPA }	**0,97260274**	0,04109589	0,95890411	Deg
{ SCTSPA }	0,622093023	0,300387597	**0,734496124**	Ppr
{ SCTSPA, SNOMEDCT }	0,418426104	0,523992322	**0,786948177**	Ppr
{ SNOMEDCT }	**1**	**1**	**1**	Deg

Considerando as recomendações apresentadas na Tabela 6.2 cada instância de *I* pode ser analisada com a finalidade de identificar a métrica mais adequada, aplicá-la e então avaliar qual dos candidatos é o mais adequado. O experimento preliminar descrito no Capítulo 5 implementou cada uma das métricas e apresentou os resultados na análise de cada instância do *corpus*. Cada conceito em análise naquele experimento utilizava o mesmo conjunto de candidatos, e indiretamente fontes desses candidatos, para cada métrica. Apenas os conceitos do contexto variam, de uma instância para outra. Sendo assim, cada conceito tem associado a ele um conjunto de fontes, oriundas dos candidatos.

Os resultados do experimento preliminar identificaram, para cada instância de cada conceito, qual ou quais métricas selecionam o sentido correto. Uma vez que a quantidade de instâncias com sentido correto por conceitos está à disposição nos resultados daquele experimento (Tabela 5.1), as recomendações de métricas pela probabilidade condicional podem ser analisadas por conceitos. Dessa forma é possível contabilizar os acertos de cada conceito em razão da métrica selecionada.

A Tabela 6.3 relaciona os conceitos do *corpus* NLM-WSD, seus respectivos conjuntos de fontes de candidatos, a métrica selecionada, o número absoluto de acertos e taxa de acertos geral em relação ao número de instâncias relevantes (i.e. coluna #totalinst da Tabela 5.1).

6.4. Resultados e Discussão

Utilizando o modelo de seleção de métricas com os dados do experimento preliminar é possível comparar os resultados com a performance dos experimentos de Agirre *et al.* [3]. Além disso, é possível mensurar os resultados parciais dos subconjuntos de conceitos propostos em Humphrey *et al.* [21] e McInnes [29].

De acordo com a Tabela 6.2, os resultados da utilização do modelo de seleção de métricas conduzem a um total de 2728 instâncias corretamente classificadas, estabelecendo uma taxa de acerto de 68,48%. Ela é 3,52%[14] melhor que a obtida com a reprodução dos experimentos com o PageRank Personalizado, onde foram classificadas 2635 instâncias corretamente, com 66,16% de acerto. O resultado é significativa, considerando os acertos exclusivos de cada proposta (teste de McNemar com uma distribuição binomial e nível de significância $p<0,001$). Na análise da performance dos 12 conceitos considerados difíceis [53], o modelo tem uma performance idêntica à do PageRank personalizado, uma vez que o modelo selecionou, para todos estes conceitos, o PageRank Personalizado. Em relação aos conceitos selecionados em Humphrey *et al*. [21] o modelo é 4,07% melhor, com 664 instâncias classificadas corretamente, contra 638 classificadas pelo PageRank Personalizado. Quando são analisados os conceitos difíceis nessa seleção, a performance é idêntica à do PageRank personalizado, pois estes conceitos são um subconjunto do conjunto dos difíceis.

Algumas heurísticas foram extraídas do relacionamento entre as métricas e os conjuntos de fontes. Quando as fontes dos candidatos forem { SNOMEDCT }, { NCI } ou { AOD e MTH }, e nenhuma outra, todas as métricas podem ser empregadas. O PageRank Personalizado é a métrica recomendada quando as fontes dos candidatos forem { CHV }, { MSH } ou { SCTSPA } e nenhuma outra. O conjunto de fontes { NDFRT, SCTSPA } determina que a métrica Degree deve ser empregada. No caso em que todas as métricas podem ser empregadas, a de menor complexidade algorítmica pode ser escolhida. Nesse caso, um total de 157 instâncias, cerca de 4,42% do total relevante, seria afetado com o emprego da métrica Degree.

Para destacar os resultados a Figura 6.5 resume todas as heurísticas identificadas pelo cálculo da probabilidade condicional. Cerca de 60% das métricas escolhidas determinaram o melhor resultado. Cinco das métricas selecionadas (cerca de 10%) levaram a um erro acima de 50% das instâncias classificadas. Em quatro desses casos (conceitos *condition*, *fit*, *inhibtion* e *sensitivity*) a métrica com melhor performance foi *Key Player Problem*, mas a escolhida foi o PageRank Personalizado. O caso restante (conceito *single*) deveria ser classificado com o PageRank Personalizado ao invés do *Key Player Problem*. Os demais erros variaram entre 1% e 32% (média 11,48% e um desvio padrão de 12,39%).

O experimento com aprendizado de máquina também foi cogitado para ser empregado na seleção de métricas. No entanto, o modelo de aprendizado extraído utiliza a fonte do conceito correto, além de outras *features* como tipos semânticos, para determinar a(s) métrica(s) mais adequada(s). Identificar o conceito correto é a etapa seguinte no processo de desambiguação (descrito na Figura 6.1) e portanto não poderia ser utilizado na etapa de seleção de métrica, conforme já expresso nesse texto.

Experimentos com a utilização dos conjuntos de fontes como *features* também foram executados. Contudo, os resultados ficaram todos abaixo dos até aqui levantados (cerca de 30% de taxa de acerto).

Uma limitação do modelo proposto é a utilização de novos conjuntos de fontes. A probabilidade condicional foi estabelecida com base nos conjuntos de fontes presentes no NLM-

14 Este valor representa a diferença percentual entre os valores absolutos de acerto de cada modelo (simples vs. híbrido).

WSD. Isso significa que, para um *corpus* com novos conceitos, haverá a necessidade da reavaliação ou adaptação das relações apresentadas neste modelo (Tabela 6.2).

Lista de conjuntos de fontes e métricas sugeridas, respectivamente:
- { AOD, MTH } → Deg
- { CHV } → Ppr
- { CHV, MTH } → Kpp
- { CHV, SCTSPA } → Ppr
- { CHV, SNOMEDCT } → Ppr
- { MSH } → Ppr
- { MSH, SCTSPA } → Ppr
- { MSH, SCTSPA, SNOMEDCT } → Ppr
- { MSH, SNOMEDCT } → Ppr
- { MTH, SCTSPA } → Kpp
- { MTH, SNOMEDCT } → Ppr
- { NCI } → Deg
- { NCI, SCTSPA } → Ppr
- { NCI, SNOMEDCT } → Kpp
- { NDFRT, SCTSPA } → Deg
- { SCTSPA } → Ppr
- { SCTSPA, SNOMEDCT } → Ppr
- { SNOMEDCT } → Deg

De forma exclusiva temos:
- se apenas { SNOMEDCT }, { NCI } ou { AOD e MTH } → todas as métricas
- se apenas { CHV }, { MSH } ou { SCTSPA } → Ppr
- se apenas { NDFRT, SCTSPA } → Deg

Figura 6.5: Heurísticas identificadas para o Modelo Híbrido

Algumas propostas para a generalização do Modelo Híbrido foram investigadas. A correlação entre atributos ou características do grafo das fontes e as métricas de seleção de sentido foi investigada. Foram avaliadas a densidade do grafo e o grau médio das arestas.

Um grafo denso é aquele que está próximo do número máximo de arestas. Para cada fonte a densidade é estabelecida pelo conjunto de vértices V (conceitos da fonte) e pelo conjunto de arestas A (relacionamentos entre os conceitos), e calculada por:

$$D = \frac{2|A|}{|V|(|V|-1)}$$

A densidade D de um grafo varia entre 0 e 1, onde zero representa um grafo sem arestas e um representa aquele totalmente conectado (cada vértice possui uma aresta para todos os outros vértices). A densidade de um conjunto de fontes foi calculada a partir da média da densidade das fontes do conjunto. Por exemplo, as densidades das fontes do conjunto { AOD, MTH } são respectivamente { 0,000959186, $1{,}42896^{-05}$ }. Portanto, a média da densidade destas fontes é 0,000486738. Esta média é então associada à métrica que de acordo com o Modelo é considerada a mais indicada, neste caso a métrica Degree. A Tabela 6.4 relaciona a densidade de cada fonte de F. A Tabela 6.5 apresenta a densidade média de cada conjunto de fontes.

Tabela 6.3: Seleção de métricas para os conceitos do NLM-WSD

Conceito	Conjunto das fontes	Selecionada	#acertos
adjustment	{ MSH, SCTSPA, SNOMEDCT }	Ppr	29
blood_pressure	{ SCTSPA }	Ppr	47
cold	{ SCTSPA, SNOMEDCT }	Ppr	24
condition	{ SCTSPA, SNOMEDCT }	Ppr	33
culture	{ NCI, SNOMEDCT }	Kpp	16
degree	{ CHV, SCTSPA }	Ppr	61
depression	{ SCTSPA }	Ppr	60
determination	{ CHV, SCTSPA }	Ppr	36
discharge	{ SCTSPA, SNOMEDCT }	Ppr	53
energy	{ CHV, SCTSPA }	Ppr	98
evaluation	{ SCTSPA }	Ppr	52
extraction	{ MTH, SCTSPA }	Kpp	82
failure	{ CHV }	Ppr	8
fat	{ NDFRT, SCTSPA }	Deg	71
fit	{ CHV, SNOMEDCT }	Ppr	2
fluid	{ SCTSPA }	Ppr	90
frequency	{ SCTSPA, SNOMEDCT }	Ppr	92
ganglion	{ NCI, SCTSPA }	Ppr	77
glucose	{ CHV, SCTSPA }	Ppr	90
growth	{ CHV, MTH }	Kpp	63
immunosuppression	{ MSH, SCTSPA }	Ppr	57
implantation	{ MSH, SNOMEDCT }	Ppr	86
inhibition	{ MSH, SCTSPA }	Ppr	14
japanese	{ SCTSPA }	Ppr	72
lead	{ SCTSPA, SNOMEDCT }	Ppr	27
man	{ SCTSPA, SNOMEDCT }	Ppr	32
mole	{ NCI, SNOMEDCT }	Kpp	83
mosaic	{ MTH, SNOMEDCT }	Ppr	77
nutrition	{ MSH, SNOMEDCT }	Ppr	29
pathology	{ MTH, SCTSPA }	Kpp	78
pressure	{ SCTSPA, SNOMEDCT }	Ppr	91
radiation	{ NCI }	Deg	61
reduction	{ SCTSPA }	Ppr	5
repair	{ SCTSPA }	Ppr	53
resistance	{ AOD, MTH }	Deg	3
scale	{ SCTSPA, SNOMEDCT }	Ppr	50
secretion	{ MSH, SCTSPA }	Ppr	99
sensitivity	{ MTH, SNOMEDCT }	Ppr	24
sex	{ CHV, SNOMEDCT }	Ppr	88
single	{ MTH, SCTSPA }	Kpp	6
strains	{ CHV, SCTSPA }	Ppr	63
support	{ SCTSPA, SNOMEDCT }	Ppr	8
surgery	{ MSH }	Ppr	98
transient	{ MSH, SNOMEDCT }	Ppr	98
transport	{ SNOMEDCT }	Deg	93
ultrasound	{ MSH, SNOMEDCT }	Ppr	84
variation	{ MSH, SCTSPA }	Ppr	80
weight	{ NCI, SNOMEDCT }	Kpp	24
white	{ CHV, SNOMEDCT }	Ppr	61
		#acertos	2728
		%acertos	68,48

Em particular, a fonte CHV é diferente das demais. Ela não possui relações entre seus conceitos[15] (não constam na tabela MRREL). Seus conceitos foram associados a outros conceitos presentes no UMLS quando da sua inserção no metatesauro. Portanto, os vértices podem ser estimados, a partir da tabela MRCONSO, mas suas arestas não.

Os intervalos de valores médios de densidade estabelecidos para cada métrica foram utilizados na elaboração da Figura 6.6. Nele é possível observar que, para fontes com densidades acima de 0,000140871 (o máximo de Ppr), é recomendável utilizar a métrica Degree. Por outro lado, não é possível identificar intervalos que caracterizem as métricas Kpp e Ppr.

Tabela 6.4: Número de arestas, vértices e densidade de cada fonte

Fonte	arestas	vértices	densidade
CHV	0	148411	0
MTH	1009807	375946	1,43E-05
MSH	2518881	316597	5,03E-05
SNOMEDCT	2732902	323027	5,24E-05
SCTSPA	2766629	315318	5,57E-05
NCI	704395	78938	0,000226089
NDFRT	414295	39581	0,000528904
AOD	121162	15895	0,000959186

Tabela 6.5: Densidade de cada fonte e a média do conjunto das fontes

Conjunto das Fontes	Fonte 1	Fonte 2	Fonte 3	Média	Métrica
{ AOD, MTH }	0,000959186	1,43E-05	0	0,000486738	Deg
{ CHV }	0	0	0	0	Ppr
{ CHV, MTH }	0	1,43E-05	0	1,43E-05	Kpp
{ CHV, SCTSPA }	0	5,57E-05	0	5,57E-05	Ppr
{ CHV, SNOMEDCT }	0	5,24E-05	0	5,24E-05	Ppr
{ MSH }	5,03E-05	0	0	5,03E-05	Ppr
{ MSH, SCTSPA }	5,03E-05	5,57E-05	0	5,30E-05	Ppr
{ MSH, SCTSPA, SNOMEDCT }	5,03E-05	5,57E-05	5,24E-05	5,28E-05	Ppr
{ MSH, SNOMEDCT }	5,03E-05	5,24E-05	0	5,13E-05	Ppr
{ MTH, SCTSPA }	1,43E-05	5,57E-05	0	3,50E-05	Kpp
{ MTH, SNOMEDCT }	1,43E-05	5,24E-05	0	3,33E-05	Ppr
{ NCI }	0,000226089	0	0	0,000226089	Deg
{ NCI, SCTSPA }	0,000226089	5,57E-05	0	0,000140871	Ppr
{ NCI, SNOMEDCT }	0,000226089	5,24E-05	0	0,000139235	Kpp
{ NDFRT, SCTSPA }	0,000528904	5,57E-05	0	0,000292278	Deg
{ SCTSPA }	5,57E-05	0	0	5,57E-05	Ppr
{ SCTSPA, SNOMEDCT }	5,57E-05	5,24E-05	0	5,40E-05	Ppr
{ SNOMEDCT }	5,24E-05	0	0	5,24E-05	Deg

15 A documentação *on-line* menciona este fato em http://www.nlm.nih.gov/research/umls/sourcereleasedocs/2011AA/CHV/relationships.html (Último acesso 10 de Fevereiro de 2013)

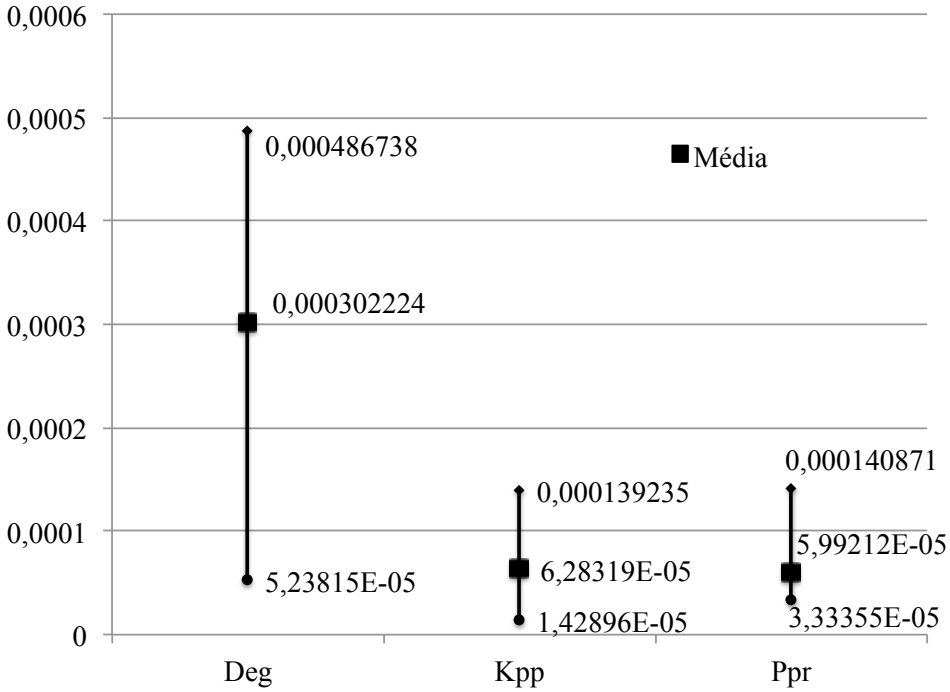

Figura 6.6: Valores máximo, mínimo e médio das médias de densidade dos conjuntos de fontes

A alternativa de avaliar o grau médio dos vértices das fontes também foi explorada. Para cada vértice de cada fonte foi contabilizado o número de arestas. O maior grau, o menor grau e grau médio de todos os vértices foram calculados para cada fonte. Esses resultados são apresentados na Tabela 6.6.

Tabela 6.6: Grau mínimo, máximo e médio de cada fonte

Fonte	grau mínimo	grau máximo	grau médio
CHV	0	0	0
MSH	1	31291	1,08691
MTH	1	3395	1,50537
NCI	1	10568	1,6121
AOD	1	83	3,17569
SCTSPA	1	80434	4,10772
SNOMEDCT	1	80910	4,61228
NDFRT	1	8149	21,75

Os conjuntos de fontes foram então avaliados pelo grau médio desses conjuntos, apresentado na Tabela 6.7. Os valores máximo, mínimo e médio das médias de densidade dos conjuntos de fontes foram analisados. Os intervalos estabelecidos estão registrados na Figura 6.7.

Os intervalos permitem identificar a relação entre algumas métricas e o grau médio das fontes. A primeira constatação é de que conjuntos de fontes com grau superior a 4,612280 arestas devem ser analisados pela métrica *Deg*. Também é possível estabelecer outros três intervalos. São eles: [0; 1,505370] que leva ao uso de *Ppr*, (1,505370; 3,112190] que leva ao emprego de *Kpp* e (3,112190; 4,612280] que leva à utilização de *Ppr*. Estes intervalos foram utilizados como forma de

selecionar a métrica de cada conceito de acordo o grau médio do conjunto das fontes. No entanto, essa estratégia resulta numa performance inferior à dos demais resultados (cerca de 58,14%).

De qualquer forma, a possibilidade de generalizar o emprego das fontes como forma de selecionar a métrica mais adequada deve ser mais amplamente investigada, o que só seria possível para além do escopo dessa tese.

Tabela 6.7: Grau médio das fontes e média geral do conjunto das fontes

Conjunto das Fontes	Fonte 1	Fonte 2	Fonte 3	Média	Métrica
{ AOD, MTH }	3,17569	1,50537	0	2,340530	Deg
{ CHV }	0	0	0	0	Ppr
{ CHV, MTH }	0	1,50537	0	1,505370	Kpp
{ CHV, SCTSPA }	0	4,10772	0	4,107720	Ppr
{ CHV, SNOMEDCT }	0	4,61228	0	4,612280	Ppr
{ MSH }	1,08691	0	0	1,086910	Ppr
{ MSH, SCTSPA }	1,08691	4,10772	0	2,597315	Ppr
{ MSH, SCTSPA, SNOMEDCT }	1,08691	4,10772	4,61228	3,268970	Ppr
{ MSH, SNOMEDCT }	1,08691	4,61228	0	2,849595	Ppr
{ MTH, SCTSPA }	1,50537	4,10772	0	2,806545	Kpp
{ MTH, SNOMEDCT }	1,50537	4,61228	0	3,058825	Ppr
{ NCI }	1,6121	0	0	1,612100	Deg
{ NCI, SCTSPA }	1,6121	4,10772	0	2,859910	Ppr
{ NCI, SNOMEDCT }	1,6121	4,61228	0	3,112190	Kpp
{ NDFRT, SCTSPA }	21,75	4,10772	0	12,928860	Deg
{ SCTSPA }	4,10772	0	0	4,107720	Ppr
{ SCTSPA, SNOMEDCT }	4,10772	4,61228	0	4,360000	Ppr
{ SNOMEDCT }	4,61228	0	0	4,612280	Deg

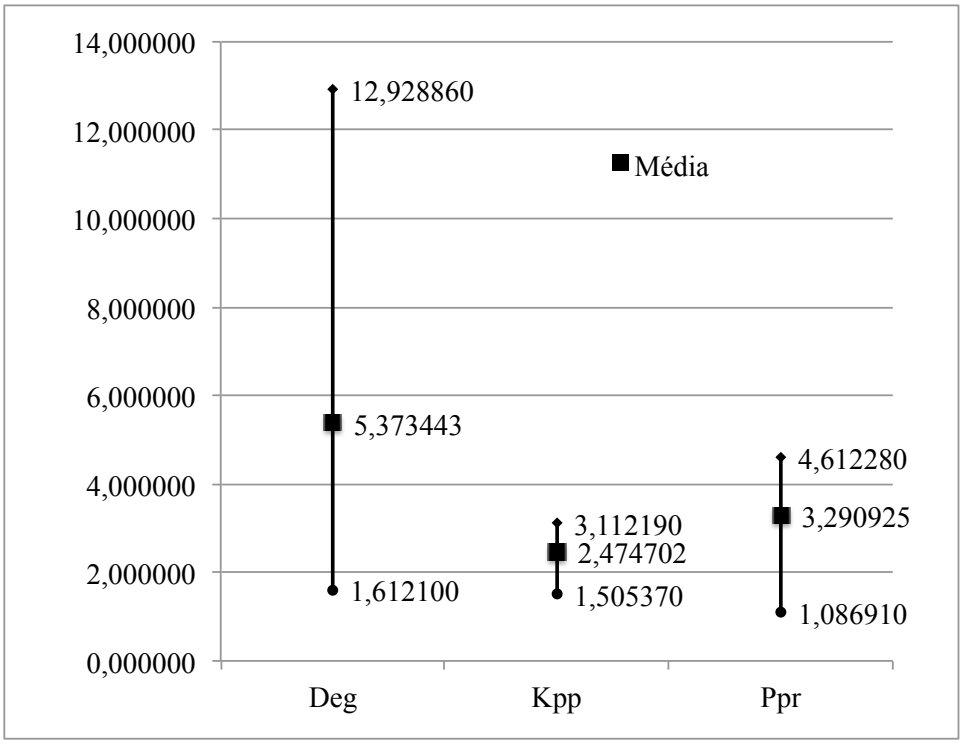

Figura 6.7: Valores máximo, mínimo e médio do grau médio dos conjuntos de fontes

7. CONCLUSÕES E TRABALHOS FUTUROS

Mesmo sendo o *Word Sense Disambiguation* uma tarefa antiga do Processamento da Linguagem Natural, é ainda um tema de pesquisa em aberto. Exploradas na forma de experimentos com domínios específicos ou de âmbito geral, as propostas para selecionar o sentido correto utilizam diferentes métodos computacionais.

7.1. O trabalho realizado

Este trabalho investigou soluções para o domínio da Biomedicina. Foram explorados um *corpus* elaborado especificamente para este domínio, fontes de conhecimento externo estruturadas e métodos para a seleção não-supervisionada do sentido correto. O principal benefício desses métodos é a sua manutenção, pois não há necessidade de treino, ou seja, novos casos de ambiguidade solucionados, para o aprendizado. O conhecimento é retirado de fontes externas que, se mantidas e aprimoradas, permitem a evolução dos resultados.

A abordagem de grafos foi escolhida, nesta pesquisa, como meio de selecionar o sentido mais adequado. Os candidatos a sentido de uma instância ambígua estão presentes numa fonte de conhecimento estruturada. Nesta área da Biomedicina, a fonte escolhida, o UMLS, é utilizada amplamente pelos pesquisadores na investigação, além de ser usada em sistemas voltados para os profissionais. O metatesauro UMLS reúne formalmente uma série de relacionamentos semânticos entre os conceitos de diferentes fontes, que são mantidas por diferentes instituições.

Para utilizar o grafo de uma fonte estruturada na seleção de um sentido é necessário escolher uma métrica. Ela serve como um meio de classificar os vértices (conceitos) de um grafo. Várias métricas foram empregadas em domínios não especializados. Degree, Key Player Problem e o PageRank Personalizado são algumas delas. No domínio da Biomedicina o PageRank Personalizado se destacou como uma das melhores métricas para seleção de sentido não-supervisionada. No entanto, o algoritmo PageRank personalizado é de maior complexidade e em domínios não especializados outras métricas se destacaram.

Este trabalho investigou e avaliou o emprego das três métricas anteriormente citadas, Degree, Key Player Problem e o PageRank Personalizado, no domínio da Biomedicina. Elas correspondem aos melhores resultados encontrados no domínio geral e específico, obtidos até então. A hipótese inicialmente levantada com este trabalho era de que as melhores métricas para o domínio não especializado também o são para o domínio especializado da Biomedicina. Elas foram implementadas e os resultados comparados. A conclusão identificou o PageRank Personalizado como a melhor escolha no contexto dos experimentos realizados. Contudo, os resultados obtidos foram analisados e revelaram a possibilidade de melhoria. Inicialmente os resultados foram avaliados no formato usual da área. Cada instância de palavra ambígua presente no *corpus* determinou a taxa de acerto geral para cada métrica. As taxas de acerto eram então comparadas umas com as outras. Na sequência, nosso estudo comparou as instâncias corretamente classificadas por cada métrica e identificou que, juntas, as métricas poderiam alcançar uma taxa de acerto maior que cada uma delas individualmente.

Levantou-se então a perspectiva da existência de um meio de identificar a métrica mais adequada para uma dada instância e assim atingir uma performance melhor do que as partes individualmente. A identificação da métrica passa pela necessidade de estabelecer *features* e uma heurística para, a partir delas, selecionar-se a métrica mais adequada. Um modelo de como

selecionar as *features,* baseado em heurísticas, foi estabelecido e denominado Modelo Híbrido. Este modelo se baseia na seleção de uma métrica a partir da probabilidade condicional entre a métrica e o conjunto das fontes dos candidatos a sentido de uma determinada instância. Cada instância ambígua foi anotada no *corpus* com um conjunto de candidatos a sentido. Estes candidatos a sentido estão associados a um conceito no UMLS e, consequentemente, a sua fonte original. A probabilidade condicional de uma métrica ocorrer quando é dado um determinado conjunto de fontes foi calculada, para cada conceito. Isso é possível por que os candidatos de cada conceito são os mesmos, para todas as suas instâncias ambíguas. Uma relação entre a métrica com maior probabilidade e o conjunto de fontes foi então construída. Utilizando este modelo e as heurísticas estabelecidas a partir dele, o par (*métrica*, *fonte*), uma nova avaliação foi executada. Nela, a decisão de qual métrica a utilizar para cada conceito foi determinada por aquele que possui a maior probabilidade de sucesso, ou seja, a maior probabilidade condicional. Os resultados revelaram um desempenho significativamente melhor do que o melhor resultado individual (PageRank Personalizado).

Além dessa contribuição, que é o ganho de 3,52%, em relação ao melhor resultado disponível, o trabalho traz outras contribuições. Apresentou de forma organizada resultados que indicaram o espaço de melhoria que se pode alcançar, ao trabalhar com múltiplas métricas. E é o primeiro, até onde sabemos, a propor, e implementar, testar e avaliar uma estratégia de articulação de diferentes métricas através de um Modelo Híbrido para o WSD em Biomedicina. O trabalho que precedeu esse esforço incluiu minuciosa análise referente a seleção de *features* e heurísticas. Os estudos sobre a generalização do emprego de fontes levaram ao aprofundamento, com a mais experimentos fazendo o uso da densidade e grau de vértices. Finalmente, o material disponível em http://www.rodrigo.goulart.nom.br/tese/ permite o acesso aos dados e resultados dos experimentos com aprendizado de máquina.

7.2. Limitações e oportunidade de aprimoramento

As limitações identificadas neste trabalho conduzem a um conjunto de oportunidades de novas pesquisas. A primeira diz respeito à limitação dos resultados a um conjunto de fontes existentes. Testes com novas métricas, novas *features*, emprego de novas versões do UMLS, além de novos *corpora*, são assuntos que devem ser explorados.

A limitação dos resultados com relação às fontes existentes, já discutida no Capitulo 6, diz respeito à possibilidade de extrair características específicas das fontes do UMLS utilizadas neste estudo. Elas poderiam ser utilizadas na generalização das probabilidades condicionais extraídas entre métricas e conjunto de fontes. O modelo passaria a representar essas heurísticas com um relacionamento entre métricas e característica(s). Duas possibilidades foram exploradas neste trabalho. A densidade e o grau médio dos grafos das fontes. Como foi discutido no capítulo anterior, nenhuma destas obteve um resultado satisfatório.

Outro ponto se refere a novas métricas que podem ser incluídas. Considerando a variedade de fontes que fazem parte do UMLS, e a sua constante atualização, novas métricas podem ser úteis nos casos em que todas as métricas aqui exploradas erraram. Neste trabalho 436 instâncias não tiveram seu sentido identificado corretamente pelas métricas avaliadas. Elas representam 10,94% das instâncias relevantes (não anotadas como *none*).

Outro ponto a destacar são as novas *features* que podem ser exploradas. Elas podem servir tanto às métricas como ao modelo de seleção de métricas. Nenhuma das métricas e nenhum dos autores aqui investigados utilizou algum dos tipos de relações especializadas presentes no UMLS. Existem 191 tipos de relações entre conceitos expressos na versão UMLS utilizada nesta pesquisa.

Utilizar subconjuntos dessas relações pode influenciar a performance das métricas. Por exemplo, cada métrica poderia fazer uso das relações do tipo QB (*qualified by*) o que determinaria relações mais qualificadas de sinonímia.

O UMLS é atualizado em média duas vezes por ano. A manutenção é decorrente das atualizações das fontes que o compõem. Neste trabalho se optou por utilizar a mesma versão do UMLS empregada nos experimentos de Agirre *et al.* [3], a 2007AB. O objetivo era reproduzir da maneira mais próxima os experimentos em [3]. No entanto, experimentos com versões novas do metatesauro não foram realizados. Um ponto importante é que o UMLS não necessariamente aumenta em quantidade de conceitos e relacionamentos, mas pode melhorar em qualidade. As fontes atualizam os conceitos e seus relacionamentos na busca de redundâncias ou erros. Por exemplo, enquanto a fonte MSH aumentou o número de conceitos da versão 2011AA para versão 2012AA, a fonte SNOMEDCT diminuiu o número de conceitos. Por outro lado, algumas não recebem atualizações com tanta frequência, como é o caso da NCI (sua última atualização foi feita na versão 2011AB). Essas mudanças influenciam diretamente a performance das métricas. Um estudo comparativo entre as versões do UMLS podem comprovar ou não essa influência. O Modelo Híbrido proposto aqui salientará ainda mais as mudanças entre versões. Uma vez que as fontes são empregadas na tomada de decisão por uma métrica, é grande a probabilidade de a qualificação das fontes influenciar a performance deste Modelo.

As pesquisas desenvolvidas para o WSD em Biomedicina fazem recorrente referência ao *corpus* NLM-WSD. Ele foi criado em 2001 utilizando os conceitos presentes no UMLS na versão 1999. No entanto, um outro *corpus* foi desenvolvido no ano de 2011, chamado MSH-WSD [23]. O *corpus* possui 203 conceitos ambíguos anotados e cerca de 37 mil instâncias ambíguas. Por se tratar de um conjunto de dados maior, ele pode trazer novas conclusões sobre como ampliar os resultados do modelo proposto.

Certamente, com o uso desse novo *corpus* no desenvolvimento das pesquisas em WSD o Modelo Híbrido também deverá ser testado e adaptado, o que não diminui a sua contribuição.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Agirre, E., and Edmonds, P.G. “Word sense disambiguation : algorithms and applications”. Springer, 2006, 388p.

[2] Agirre, E., and Soroa, A. “Personalizing PageRank for word sense disambiguation”. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009, pp. 2889–2896.

[3] Agirre, E., Soroa, A., and Stevenson, M. “Graph-based Word Sense Disambiguation of biomedical documents”, Bioinformatics, vol. 26-22, 2010, pp. 2889-2896.

[4] Aronson, A.R. “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.”. In: Proceedings of American Medical Informatics Association (AMIA) Symposium, 2001, pp. 17-21.

[5] Balmin, A., Hristidis, V., and Papakonstantinou, Y. “Objectrank: authority-based keyword search in databases”. In: Proceedings of the Thirtieth international conference on Very large data bases, 2004, pp 564-575.

[6] Bollen, J., Rodriguez, M., and Van de Sompel, H. “MESUR: usage-based metrics of scholarly impact”. In: Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries, 2007, pp. 474-474.

[7] Borgatti, S.P. “Identifying sets of key players in a social network”, Comput. Math. Organ. Theory, vol. 12, 2006, pp. 21-34.

[8] Brin, S., and Page, L. “The anatomy of a large-scale hypertextual Web search engine”, Computer Networks and ISDN Systems, vol. 30, Janeiro-Julho 1998, pp. 107-117.

[9] Browne, A.C., McCray, A.T., and Srinivasan, S. “The SPECIALIST Lexicon”, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 2000, pp. 96.

[10] Bruce, R.F., and Wiebe, J.M. “Recognizing subjectivity: a case study in manual tagging”, Natural Language Engineering, vol. 5(2), 1999, pp. 187-205.

[11] Chakrabarti, S., Berg, M.v.d., and Dom, B. “Focused crawling: a new approach to topic-specific Web resource discovery”, Computer Networks, vol. 31, 1999, pp. 1623-1640.

[12] Deuter, M., and Lea, D. “Oxford collocations dictionary: for students of English”, Oxford University Press, 2002, pp 1075-1086.

[13] Esuli, A., and Sebastiani, F. “PageRanking WordNet Synsets: An Application to Opinion Mining”. In: Proc. 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp 424-431.

[14] Feng, D., Shaw, E., Kim, J., and Hovy, E. “Learning to detect conversation focus of threaded discussions”. In: Proc. Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2006, pp 208-215.

[15] FLEISS, J. “Measuring nominal scale agreement among many raters.”, Psychological Bulletin, vol. 76(5), Novembro 1971, pp 378-382.

[16] Freeman, L.C. “Centrality in social networks conceptual clarification”, Social networks, vol. 33, 1979, pp. 215-239.

[17] Galley, M., and McKeown, K. “Improving Word Sense Disambiguation in Lexical Chaining”. In: Procedings of the *18th International Joint Conference on Artificial Intelligence (IJCAI)*, 2003, 3p.

[18] Gruber, T.R. “A translation approach to portable ontology specifications”, Knowledge Acquisition, vol. 5, Fevereiro 1993, pp. 199-220.

[19] Gyngyi, Z., Garcia-Molina, H., and Pedersen, J. “Combating web spam with trustrank”. In: Proceedings of the Thirtieth international conference on Very large data bases, 2004, pp 576-587 .

[20] Haveliwala, T.H. “Topic-sensitive PageRank”. In: Proc. Proceedings of the 11th international conference on World Wide Web, Honolulu, 2002, pp 517-526.

[21] Humphrey, S.M., Rogers, W.J., Kilicoglu, H., Demner-Fushman, D., and Rindflesch, T.C. “Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment”, Journal of the American Society for Information Science and Technology, vol. 57, Janeiro 2006, pp. 96-113.

[22] Humphreys, B., Lindberg, D., Schoolman, H., and Barnett, G. “The Unified Medical Language System: An informatics research collaboration”, Journal of the American Medical Informatics Association, vol. 5, 1998, pp. 1-11.

[23] Jimeno-Yepes, A., McInnes, B., and Aronson, A. “Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation”, BMC Bioinformatics, vol. 12(1), 2011, 14p.

[24] Joshi, M., Pedersen, T., and Maclin, R. “A comparative study of support vector machines applied to the supervised word sense disambiguation problem in the medical domain”, In: Indian International conference on Artificial Intelligence, 2005, 20p.

[25] Jurafsky, D., and Martin, J.H. “Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition”, Pearson Prentice Hall, 2009, 988p.

[26] Kleinberg, J. “Authoritative sources in a hyperlinked environment”, Journal of the ACM, 46, Maio 1999, pp. 604-632.

[27] Lesk, M. “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone”. In: Proc. Proceedings of the 5th annual international conference on Systems documentation, 1986, pp 24-26.

[28] Liu, H., Teller, V., and Friedman, C. “A multi-aspect comparison study of supervised word sense disambiguation”, Journal of the American Medical Informatics Association, vol. 11, 2008, 11, pp. 320-331.

[29] McInnes, B.T. “An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and Medline”. In: Proceeding HLT-SRWS '08 Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop, 2008, pp. 49-54.

[30] McInnes, B.T. “An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and Medline”. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop, 2008, pp. 49-54.

[31] Miller, G.A. “Wordnet - a Lexical Database for English”, Communications of the ACM, vol. 38, Novembro 1995, pp. 39-41.

[32] Miller, G.A., Leacock, C., Tengi, R., and Bunker, R.T. “A semantic concordance”. In: Proceedings of the workshop on Human Language Technology, 1993, pp 303-308.

[33] Navigli, R. “Semi-Automatic Extension of Large-Scale Linguistic Knowledge Bases”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, Julho 2005, pp. 548-553.

[34] Navigli, R.: “Word Sense Disambiguation: A Survey”, ACM Computing Surveys, vol. 41, 2009, 62p.

[35] Navigli, R., and Lapata, M. “An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, Abril 2010, pp. 678-692.

[36] Navigli, R., and Lapata, M. “Graph connectivity measures for unsupervised word sense disambiguation”. In: Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI) 2007, pp. 1683–1688.

[37] Navigli, R., and Velardi, P. “Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites”, Computational Linguistic, vol. 30, Fevereiro 2004, pp. 151-179.

[38] Navigli, R., and Velardi, P. “Structural semantic interconnections: A knowledge-based approach to word sense disambiguation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, Julho 2005, pp. 1075-1086.

[39] Newman, M.E.J. “The Structure and Function of Complex Networks”, Society for Industrial and Applied Mathematics (SIAM) Review, vol. 45, 2003, pp. 167-256.

[40] National Library of Medicine (US). “UMLS® Reference Manual”, Capturado em http://www.ncbi.nlm.nih.gov/books/NBK9676/, Julho de 2011.

[41] Otterbacher, J., G, ne, Erkan, and Radev, D.R. “Using random walks for question-focused sentence retrieval”. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp 915-922.

[42] Page, L., Brin, S., Motwani, R., and Winograd, T. “The PageRank Citation Ranking: Bringing Order to the Web”, In: “Book The PageRank Citation Ranking: Bringing Order to the Web”, Stanford University, 1999, 17p.

[43] Paul, D.B., and Baker, J.M. “The design for the wall street journal-based CSR *corpus*”. In: Proceedings of the workshop on Speech and Natural Language, 1992, pp 357-362.

[44] Pradhan, S.S., Loper, E., Dligach, D., and Palmer, M. “SemEval-2007 task 17: English lexical sample, SRL and all words”. In: Proceedings of the 4th International Workshop on Semantic Evaluations, 2007, pp 87-92.

[45] Pustejovsky, J. “The generative lexicon”. MIT Press, 1995, 312p.

[46] Sanderson, M. “Word sense disambiguation and information retrieval”. In: Proceedings of the 17th annual international ACM Special Interest Group on Information Retrieval (SIGIR) conference on Research and development in information retrieval,1994, pp. 142-151.

[47] Savova, G.K., Coden, A.R., Sominsky, I.L., Johnson, R., Ogren, P.V., Groen, P.C.d., and Chute, C.G. “Word sense disambiguation across two domains: Biomedical literature and clinical notes”, Journal of Biomedical Informatics, vol. 41, Junho 2008, pp. 1088-1100.

[48] Sinha, R., and Mihalcea, R. “Unsupervised Graph-basedWord Sense Disambiguation Using Measures of Word Semantic Similarity”. In: IEEE International Conference on Semantic Computing, 2007, pp. 363-369.

[49] Snyder, B., and Palmer, M. “The English all-words task”. In: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004, pp. 41-43.

[50] Summers, D. “Longman language activator: helps you write and speak natural English”. Longman, 2008, 1530p.

[51] Tsatsaronis, G., Vazirgiannis, M., and Androutsopoulos, I. “Word sense disambiguation with spreading activation networks generated from thesauri”. In: Proceedings of the 20th international joint conference on Artifical intelligence (IJCAI), 2007, pp. 1725-1730.

[52] Wan, X., and Yang, J. “Improved affinity graph based multi-document summarization”. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2006, pp 181-184.

[53] Weeber, M., Mork, J.G., and Aronson, A.R. “Developing a test collection for biomedical word sense disambiguation.”, In: Proceedings of American Medical Informatics Association (AMIA) Symposium, 2001, pp. 746-750.

APÊNDICE A

Tabela A.1: Lista de Conceitos vs. Fontes de Candidatos

Conceito	Conjunto das fontes F_i
adjustment	{ MSH, SCTSPA, SNOMEDCT }
blood_pressure	{ SCTSPA }
cold	{ SCTSPA, SNOMEDCT }
condition	{ SCTSPA, SNOMEDCT }
culture	{ NCI, SNOMEDCT }
degree	{ CHV, SCTSPA }
depression	{ SCTSPA }
determination	{ CHV, SCTSPA }
discharge	{ SCTSPA, SNOMEDCT }
energy	{ CHV, SCTSPA }
evaluation	{ SCTSPA }
extraction	{ MTH, SCTSPA }
failure	{ CHV }
fat	{ NDFRT, SCTSPA }
fit	{ CHV, SNOMEDCT }
fluid	{ SCTSPA }
frequency	{ SCTSPA, SNOMEDCT }
ganglion	{ NCI, SCTSPA }
glucose	{ CHV, SCTSPA }
growth	{ CHV, MTH }
immunosuppression	{ MSH, SCTSPA }
implantation	{ MSH, SNOMEDCT }
inhibition	{ MSH, SCTSPA }
japanese	{ SCTSPA }
lead	{ SCTSPA, SNOMEDCT }
man	{ SCTSPA, SNOMEDCT }
mole	{ NCI, SNOMEDCT }
mosaic	{ MTH, SNOMEDCT }
nutrition	{ MSH, SNOMEDCT }
pathology	{ MTH, SCTSPA }
pressure	{ SCTSPA, SNOMEDCT }
radiation	{ NCI }
reduction	{ SCTSPA }
repair	{ SCTSPA }
resistance	{ AOD, MTH }
scale	{ SCTSPA, SNOMEDCT }
secretion	{ MSH, SCTSPA }
sensitivity	{ MTH, SNOMEDCT }
sex	{ CHV, SNOMEDCT }
single	{ MTH, SCTSPA }
strains	{ CHV, SCTSPA }
support	{ SCTSPA, SNOMEDCT }
surgery	{ MSH }
transient	{ MSH, SNOMEDCT }
transport	{ SNOMEDCT }
ultrasound	{ MSH, SNOMEDCT }
variation	{ MSH, SCTSPA }
weight	{ NCI, SNOMEDCT }
white	{ CHV, SNOMEDCT }

APÊNDICE B

Tabela B.1: Relação de frequência de instâncias na interseção entre métricas (discriminadas as ocorrências em que subconjuntos de métricas identificam o sentido correto) e conjuntos de fontes

|(Fontes ∩ Métrica)|	Deg	DegKpp	DegKppPpr	DegPpr	Kpp	KppPpr	Ppr	Total
{ AOD, MTH }	0	0	3	0	0	0	0	3
{ CHV }	0	0	4	0	2	0	4	10
{ CHV, MTH }	0	0	0	37	63	0	0	100
{ CHV, SCTSPA }	2	4	43	140	59	76	89	413
{ CHV, SNOMEDCT }	18	7	30	41	23	31	49	199
{ MSH }	0	0	19	70	1	1	8	99
{ MSH, SCTSPA }	15	4	6	152	124	85	7	393
{ MSH, SCTSPA, SNOMEDCT }	1	0	1	11	14	0	17	44
{ MSH, SNOMEDCT }	14	10	6	210	35	10	71	356
{ MTH, SCTSPA }	4	60	27	17	60	19	89	276
{ MTH, SNOMEDCT }	14	0	0	40	32	38	23	147
{ NCI }	0	0	61	0	0	0	0	61
{ NCI, SCTSPA }	0	2	8	13	19	42	14	98
{ NCI, SNOMEDCT }	1	1	24	10	73	25	11	145
{ NDFRT, SCTSPA }	1	0	1	69	2	0	0	73
{ SCTSPA }	38	18	46	219	81	10	104	516
{ SCTSPA, SNOMEDCT }	21	2	62	133	88	121	94	521
{ SNOMEDCT }	0	0	93	0	0	0	0	93
							Espaço amostral -> |*I*|	3547

Tabela B.2: Relação de frequência de instâncias na interseção entre métricas (valor total pode métrica) e conjuntos de fontes, probabilidade da interseção

|(Fontes ∩ Métrica)|	$\lvert Fontes \cap Métrica \rvert$			$P(Fontes \cap Métrica)$		
	Deg	DegKpp	DegKppPpr	DegPpr	Kpp	KppPpr
{ AOD, MTH }	3	3	3	0,000845785	0,000845785	0,000845785
{ CHV }	4	6	8	0,001127714	0,00169157	0,002255427
{ CHV, MTH }	37	63	37	0,01043135	0,017761489	0,01043135
{ CHV, SCTSPA }	189	182	348	0,053284466	0,051310967	0,09811108
{ CHV, SNOMEDCT }	96	91	151	0,027065125	0,025655484	0,042571187
{ MSH }	89	21	98	0,025091627	0,005920496	0,027628982
{ MSH, SCTSPA }	177	219	250	0,049901325	0,061742317	0,070482098
{ MSH, SCTSPA, SNOMEDCT }	13	15	29	0,003665069	0,004228926	0,008175923
{ MSH, SNOMEDCT }	240	61	297	0,067662814	0,017197632	0,083732732
{ MTH, SCTSPA }	108	166	152	0,030448266	0,046800113	0,042853115
{ MTH, SNOMEDCT }	54	70	101	0,015224133	0,019734987	0,028474767
{ NCI }	61	61	61	0,017197632	0,017197632	0,017197632
{ NCI, SCTSPA }	23	71	77	0,006484353	0,020016916	0,021708486
{ NCI, SNOMEDCT }	36	123	70	0,010149422	0,034677192	0,019734987
{ NDFRT, SCTSPA }	71	3	70	0,020016916	0,000845785	0,019734987
{ SCTSPA }	321	155	379	0,090499013	0,0436989	0,10685086
{ SCTSPA, SNOMEDCT }	218	273	410	0,061460389	0,076966451	0,11559064
{ SNOMEDCT }	93	93	93	0,02621934	0,02621934	0,02621934

Tabela B.3: Probabilidade de cada conjunto das fontes

Conjunto das Fontes	$P(Fontes)$
{ AOD, MTH }	0,000845785
{ CHV }	0,002819284
{ CHV, MTH }	0,028192839
{ CHV, SCTSPA }	0,116436425
{ CHV, SNOMEDCT }	0,05610375
{ MSH }	0,027910911
{ MSH, SCTSPA }	0,110797857
{ MSH, SCTSPA, SNOMEDCT }	0,012404849
{ MSH, SNOMEDCT }	0,100366507
{ MTH, SCTSPA }	0,077812236
{ MTH, SNOMEDCT }	0,041443473
{ NCI }	0,017197632
{ NCI, SCTSPA }	0,027628982
{ NCI, SNOMEDCT }	0,040879617
{ NDFRT, SCTSPA }	0,020580772
{ SCTSPA }	0,145475049
{ SCTSPA, SNOMEDCT }	0,146884691
{ SNOMEDCT }	0,02621934
$\sum P(Fontes)$	1

ANEXO A

A lista de *stop phrases* da Figura A.1 corresponde a um conjunto de palavras utilizado pelo pacote de programas UKB (http://ixa2.si.ehu.es/ukb/). Estes programas foram utilizados na reprodução dos experimentos de Agirre *et al*. [3] neste trabalho. A lista é referenciada na Seção 4.3, página 36, e no Capítulo 5.

not
BACKGROUND
METHODS
RESULTS
CONCLUSIONS
OBJECTIVE
PURPOSE
MATERIALS
DESIGN

Figura A.1: Lista de *stop phrases*