Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Ligação de Entidades: uma nova abordagem para ligação de Conceitos Concretos com entidades Wiki utilizando Modelos de Espaço Vetorial

Lucas Borges Monteiro

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Orientador
Prof. Dr. Li Weigang

Brasília
2015

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Mestrado em Informática

Coordenadora: Prof.ª Dr.ª Alba C. M. A. Melo

Banca examinadora composta por:

Prof. Dr. Li Weigang (Orientador) — CIC/UnB
Prof. Dr. Jorge Rady de Almeida Júnior — POLI/USP
Prof. Dr. Mauricio Ayala Rincón — CIC/UnB
Prof. Dr. André Costa Drummond — CIC/UnB

CIP — Catalogação Internacional na Publicação

Monteiro, Lucas Borges.

Ligação de Entidades: uma nova abordagem para ligação de Conceitos Concretos com entidades Wiki utilizando Modelos de Espaço Vetorial / Lucas Borges Monteiro. Brasília : UnB, 2015.
97 p. : il. ; 29,5 cm.

Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2015.

1. Ligação de entidades, 2. Modelo de Espaço Vetorial, 3. Wikificação, 4. Conceitos Concretos

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília–DF — Brasil

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Ligação de Entidades: uma nova abordagem para ligação de Conceitos Concretos com entidades Wiki utilizando Modelos de Espaço Vetorial

Lucas Borges Monteiro

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Prof. Dr. Li Weigang (Orientador)
CIC/UnB

Prof. Dr. Jorge Rady de Almeida Júnior
POLI/USP

Prof. Dr. Mauricio Ayala Rincón
CIC/UnB

Prof. Dr. André Costa Drummond
CIC/UnB

Prof.ª Dr.ª Alba C. M. A. Melo
Coordenadora do Mestrado em Informática

Brasília, 20 de agosto de 2015

LUCAS BORGES MONTEIRO

LIGAÇÃO DE ENTIDADES: UMA NOVA ABORDAGEM PARA LIGAÇÃO DE CONCEITOS CONCRETOS COM ENTIDADES WIKI UTILIZANDO MEV

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Curso de Pós-graduação em Informática da Universidade de Brasília, pela Comissão formada pelos professores:

Orientador: Prof. Dr. Prof. Dr. Li Weigang
(CIC/UnB)

Prof. Dr. Maurício Ayala-Rincon
(CIC/UnB)

Prof. Dr. Jorge Rady de Almeida Júnior
(POLI/USP)

Vista e permitida a impressão.
Brasília, 20 de agosto de 2015.

Prof.ª Dr.ª Alba Cristina Magalhães Alves de Melo
Programa de Pós-Graduação em Informática
Departamento de Ciência da Computação
Universidade de Brasília

Dedicatória

Este trabalho é dedicado à toda minha família.

Agradecimentos

Em primeiro lugar agradeço a Deus.

Agradeço especialmente aos meus familiares: minha esposa Leiliane, por estar ao meu lado durante o mestrado, demonstrando apoio e paciência em todos os momentos; minha mãe Regina, por ensinar desde cedo a importância dos estudos; minha irmã Paula, pelas experiências que me transmitiu, vivenciadas durante sua trajetória acadêmica; e meu pai Paulo e meu irmão Daniel, por se fazerem sempre presentes em meu coração.

Agradeço ao meu orientador prof. Dr. Li Weigang por acreditar em mim e me acompanhar nesses dois anos de mestrado.

E finalmente, agradeço aos demais professores e funcionários do programa de pós-graduação em Informática da Universidade de Brasília.

Resumo

Ligação de Entidades (LE) é um importante tópico de pesquisa com diversas aplicações web. Apesar do crescente interesse o foco ainda tem sido a identificação de nomes próprios, isto é, pessoas, organizações, lugares, unidades de medida, etc. O principal desafio aqui é encontrar conceitos concretos (sentenças sem classe de entidade pré-definida) em textos da web conectando-os às respectivas páginas da Wikipédia.

Este trabalho apresenta uma nova abordagem para ligar conceitos concretos obtidos de textos em Inglês com entidades Wiki, neste trabalho representadas por páginas da Wikipédia, utilizando classificação gramatical (*part-of-speech*) para detectar conceitos concretos e Modelos de Espaço Vetorial (MEV) para realizar a desambiguação das entidades Wiki selecionadas da base.

A solução, denominada UnBWiki VSM, foi implementada em Java, por meio da IDE Eclipse, com banco de dados MySQL onde a base de entidades foi armazenada.

O framework proposto foi ajustado para trabalhar com uma base de Wikilinks, referências para páginas da Wikipédia extraídas de diferentes páginas da web, contendo por volta de 2,8 milhões de entidades e 18 milhões de palavras, e obteve *recall* 34,2% superior ao obtido pela metodologia existente que utilizou os mesmos dados/entidades. Como estudo de caso, textos sobre a História da Família Real Britânica extraídos da web foram analisados manualmente, e o *recall* de 73,5% obtido pela ferramenta UnBWiki VSM foi ainda maior do que o verificado na comparação com o estado da arte.

Palavras-chave: Ligação de entidades, Modelo de Espaço Vetorial, Wikificação, Conceitos Concretos

Abstract

Entity Linking (EL) is an important research topic with several web applications. Despite the growing interest the focus also has been on the identification of proper names, i.e, people, organizations, places, units of measure, and others. The main challenge here is to find concrete concepts (sentences without predefined entity class) on web texts by linking them to their respective pages of Wikipedia.

This paper presents a new approach to connect concrete concepts taken from texts in English with Wiki entities, in this work represented by the Wikipedia pages, using classification *part-of-speech* to detect concrete concepts and Vector Space Models (VSM) to perform the disambiguation of entities selected from Wiki base.

The solution, called UnBWiki VSM, was implemented in Java using the Eclipse IDE with MySQL database where the base of entities was stored.

The proposed framework was adjusted to work with a Wikilinks database, references to Wikipedia pages drawn from different web pages, containing approximately 2.8 million entities and 18 million words, and obtained *recall* 34.2% higher than the existing methodology that used the same data/entities. As a case study, Royal Family History texts extracted from the web were analyzed manually, and the *recall* of 73.5% obtained by UnBWiki VSM tool was greater than that observed in comparison with the state of the art.

Keywords: Entity linking, Vector Space Model, Wikification, Concrete Concepts

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

API Interface de Programação de Aplicações (*Application Programming Interface*)

ASL Análise Semântica Latente

CC Conceitos concretos

CIT *International Conference on Computer and Information Technology*

D2W Desambiguação com a Wikipédia (*Disambiguate to Wikipedia*)

EI Extração de Informação

EL Ligação de Entidades (*Entity Linking*)

EN Entidade Nomeada

HTTP Protocolo de Transferência de Hipertexto (*Hypertext Transfer Protocol*)

LE Ligação de Entidades

LEN Ligação de Entidades Nomeadas

MEV Modelo de Espaço Vetorial

MVS Máquinas de Vetores de Suporte

NER Reconhecimento de Entidades Nomeadas (*Named Entity Recognition*)

NP Nomes próprios

PLN Processamento de Linguagem Natural

POS Análise gramatical (*part-of-speech*)

Sa2W Anotação pontuada com a Wikipédia (*Scored-annotate to Wikipedia*)

TF-IDF Funções frequência de termos versus inverso da frequência no documento

URL Localizador Padrão de Recursos (*Uniform Resource Locator*)

VSM Modelo de Espaço Vetorial (*Vector Space Model*)

Wise *International Conference on Web Information System Engineering*

WSD Desambiguação de sentidos de palavras (*Word Sense Disambiguation*)

XML Linguagem de Marcação Extensível (*eXtensible Markup Language*)

Capítulo 1

Introdução

1.1 Contextualização

O aumento da disponibilidade de textos em formato digital, tais como notícias, livros, trabalhos acadêmicos, enciclopédias digitais, dentre outros, é um dos principais motivos para o crescente número de pessoas que realizam buscas na Internet. Estudos apontam que 92% dos usuários da Internet consideram a rede mundial de computadores um lugar propício para a busca por informações [30].

Tratar manualmente esse grande volume de informação, muitas das vezes em diferentes idiomas, é uma tarefa praticamente impossível e o uso de ferramentas automáticas de extração de informação tornou-se inevitável. Uma tarefa específica nesse processo de tratamento automático de textos envolve identificar sentenças em textos sem formato pré-definido e ligar tais termos com sua respectiva entidade, em uma base de conhecimentos específica. Essa tarefa recebe o nome de ligação de entidades (LE).

1.2 Motivação

O uso comercial de extratores/anotadores tem aumentado gradativamente e a precisão desses sistemas é um importante fator de sucesso para qualquer ferramenta. A identificação de nomes próprios apresenta menos complexidade uma vez que menções formadas por nomes próprios são mais bem sintaticamente definidas. Porém, identificar conceitos concretos é uma tarefa mais desafiadora, visto que sofre maior influência de regras gramaticais mais específicas do que as envolvidas na identificação de nomes próprios, regras estas que muitas vezes podem ser categoricamente diferentes entre idiomas distintos.

Existe na literatura uma grande variedade de ferramentas que se propõem a realizar ligação de entidades. Normalmente, essas ferramentas aplicam-se a um domínio específico ou restringem-se a um pequeno grupo de classes de palavras, tais como personalidades,

lugares ou eventos, lançando mão de mecanismos que permitem identificar corretamente termos que se enquadram em uma dessas classes.

Alguns dos estudos mais recentes têm adotado as páginas da Wikipédia e suas respectivas menções como entidades, isto porque a Wikipédia oferece um extenso catálogo atualizado diariamente, com estrutura bem definida, que aborda os mais variados tópicos cotidianos, e está disponível em diferentes idiomas. Os estudos em processamento de linguagem natural envolvendo a Wikipédia formam um ramo bem específico de pesquisa, e os problemas nessa área podem ser classificados basicamente em seis grupos [7]:

- *Disambiguate to Wikipedia (D2W)*: compreende a escolha da entidade mais pertinente para cada menção identificada;
- *Annotate to Wikipedia (A2W)*: trata da identificação das menções mais relevantes e das respectivas entidades;
- *Scored-annotate to Wikipedia (Sa2W)*: consiste em um problema semelhante ao A2W mas que envolve a atribuição de pontuação no processo de escolha;
- *Concepts to Wikipedia (C2W)*: consiste na identificação de tópicos a partir das entidades selecionadas;
- *Scored-concepts to Wikipedia (Sc2W)*: trata-se problema semelhante ao C2W mas que envolve a atribuição de pontuação no processo de escolha; e
- *Ranked-concepts to Wikipedia (Rc2W)*: compreende a classificação das entidades selecionadas de acordo com sua relevância para cada tópico do texto.

Nesse universo de problemas surgem duas figuras centrais de busca que possuem papel fundamental no contexto dos textos: nomes próprios (NP) e conceitos concretos (CC). A identificação de NP – personalidades, lugares, siglas, instituições, dentre outras – é um problema clássico há tempos explorado pelas pesquisas em ligação de entidades. Por sua vez, a identificação de CC está presente de maneira variada em alguns estudos, sendo que a própria definição de CC carrega consigo certa ambiguidade, podendo ser incluída em uma categoria de problema em separado.

A ligação de entidades compostas por classes de palavras não triviais - os conceitos concretos - é um desafio pouco investigado até o momento, e o surgimento de ferramentas com essa capacidade tende a aumentar sobremaneira a eficácia de anotadores automáticos e extratores de informação.

1.3 Definição do Problema

O intuito deste trabalho é apresentar uma nova abordagem para a realização de ligação de entidades envolvendo conceitos concretos com entidades Wiki. Esse objetivo será alcançado por meio da construção de uma ferramenta que possibilite tratar textos passados como entrada visando a identificação de NPs e CCs que serão ligados às respectivas entidades que, neste trabalho, serão representadas por páginas da Wikipédia.

O desafio é: como realizar ligação de entidades tanto de NPs como de CCs de maneira automática e obter *recall* satisfatório. Essa dissertação apresenta uma abordagem utilizando Modelos de Espaço Vetorial (MEV) para tratar esse desafio.

A relevância desse trabalho está na identificação de métodos que permitam concluir quando sequências de palavras configuram um determinado conceito concreto, obtendo resultados satisfatórios no processo de seleção da entidade mais adequada aos conceitos concretos identificados.

Apesar de já existirem ferramentas de ligação de entidades em Inglês, com índices de acertos satisfatórios, pouco se explorou a relevância da ligação de entidades envolvendo CC. Além disso, com relação ao idioma Português (brasileiro), existem poucos estudos sistemáticos relacionados à ligação de entidades, e esse trabalho pretende contribuir para a consolidação dos estudos dessa área no Brasil.

1.4 Metodologia da Pesquisa

Esta seção explica o que consiste cada uma das três etapas da ferramenta construída para aplicação da metodologia proposta.

1.4.1 Etapa 1: Criação da Base de Entidades

A etapa de criação da base de entidades consiste no tratamento das bases da Wikipédia disponibilizadas *offline* de modo a formar um banco de dados normalizado de entidades e respectivas menções, entidades que serão ligadas posteriormente aos NPs e CCs extraídos dos textos.

Nesta etapa foi desenvolvida uma ferramenta para receber os arquivos da Wikipédia como entrada, identificando para cada artigo as informações título, link (URL) e assunto, características selecionadas como parâmetros para a análise da similaridade dos artigos com as sentenças extraídas dos textos.

1.4.2 Etapa 2: Extração de Nomes Próprios e de Conceitos Concretos

Na segunda etapa o objetivo é tratar o conjunto de textos passados como entrada com vistas a identificar nomes próprios e possíveis conceitos concretos relevantes para a compreenção de cada texto. Apesar do foco ser a identificação de CC, a identificação de NP também foi mantida no escopo de análise da ferramenta. O resultado dessa etapa é a produção, para cada texto, de uma lista de NPs e CCs que dizem muito sobre os textos, e que serão alvo do processo de ligação de entidades que será realizado na próxima etapa.

A identificação de NP e CC será efetuada com o auxílio de ferramenta de análise gramatical (*part-of-speech*) contextualizada em detalhe nos capítulos seguintes.

1.4.3 Etapa 3: Ligação das Entidades com a Wikipédia

Por fim, na terceira etapa os NPs e CCs identificados na etapa anterior serão ligados às respectivas entidades contidas no banco de dados formado por páginas da Wikipédia. Para cada NP e CC extraído do texto, uma lista de entidades candidatas é formada. Tais entidades são representadas em um Modelo de Espaço Vetorial para auxiliar na seleção da entidade com significado mais semelhante ao do NP e CC tratado no momento.

Os Modelos de Espaço Vetorial tem como intenção representar cada documento em uma coleção como um ponto em um espaço (um vetor em um espaço vetorial). Pontos que estão próximos neste espaço são semanticamente similares e pontos que estão distantes são semanticamente distantes. A pesquisa que se deseja fazer, na forma de uma consulta, é também representada como um ponto no mesmo espaço em que os documentos são inseridos (a consulta é um considerada uma espécie de pseudo-documento). Os documentos são, então, classificados em ordem crescente de distância (ordem de diminuição da semelhança semântica) a partir da consulta. Por este motivo, a representação MEV foi adotada na seleção das entidades mais semanticamente semelhante aos NPs e CCs extraídos dos textos.

1.5 Objetivo Geral

O objetivo geral desta dissertação é apresentar uma nova abordagem para a realização de ligação de entidades de nomes próprios e conceitos concretos.

1.5.1 Objetivos Específicos

São objetivos específicos:

- Apresentar uma metodologia para realização de ligação de entidades utilizando representação MEV.
- Apresentar uma metodologia para a identificação de NPs e CCs nos textos.
- Analisar o processamento de diversos textos utilizando as metodologias propostas.

1.6 Organização da Dissertação

O capítulo 2 é dedicado aos Modelos de Espaço Vetorial. As seções e subseções apresentam os principais conceitos relacionados aos MEV bem como os métodos de utilização, representação matemática e etapas de pré-processamento necessárias. Por fim, são citadas as principais utilizações dos MEV.

O capítulo 3 apresenta os principais conceitos relacionados à ligação de entidades, descrevendo as classes de problemas, as principais bases de conhecimento utilizadas, as classes de entidades e os maiores desafios investigados atualmente.

O capítulo 4 descreve as pesquisas mais recentes envolvendo ligação de entidades, indicando as classes de problemas investigadas e as abordagens adotadas. Além disso, são apresentadas semelhanças e diferenças entre o estado da arte e a metodologia proposta neste trabalho.

O capítulo 5 apresenta a metodologia proposta para a realização de ligação de entidades envolvendo nomes próprios e conceitos concretos utilizando Modelos de Espaço Vetorial.

No capítulo 6 são apresentados os estudos de caso realizados para avaliação da ferramenta. O primeiro deles compreende a análise dos textos fornecidos na 14ª edição da *International Conference on Web Information System Engineering* (Wise 2013[1]) e a comparação com os resultados obtidos na época da realização do evento. O segundo estudo de caso compreende a avaliação da ferramenta no processamento de textos relacionados à história da família real britânica (*Royaltree*[2]), analisados previamente de maneira manual.

Por fim, no capítulo 7 são apresentadas as considerações finais sobre o presente trabalho e os direcionamentos para o tema abordado.

[1] http://wise2013.njue.edu.cn/wise2013challenge.html
[2] http://www.britroyals.com/royaltree.html

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentados os conceitos relacionados aos Modelos de Espaço Vetorial, teoria fundamental empregada na metodologia ora proposta. A seção 2.1 dedica-se à introdução da metodologia. A seção 2.2 trata da motivação do uso dos Modelos de Espaço Vetorial. Já a seção 2.3 tem como objetivo descrever como os Modelos de Espaço Vetorial são empregados. A funtamentação linguística e a fundamentação matemática da teoria são apresentadas nas seções 2.4 e 2.5, respectivamente. Na seção 2.6 são apresentadas as principais aplicações envolvendo Modelos de Espaço Vetorial. A seção 2.7 contextualiza o uso das métricas *precisão* e *recall*. Por fim, a seção 2.8 traz um resumo do capítulo.

2.1 Modelos de Espaço Vetorial

Um dos maiores obstáculos para se fazer pleno uso do poder dos computadores é que eles atualmente compreendem pouco do significado da linguagem humana. Esse desafio impulsiona o impacto transformador que tecnologias semânticas mais profundas têm tido recentemente. Os Modelos de Espaço Vetorial (MEV) integram esse grupo de tecnologias semânticas transformadoras [54].

A semântica aqui é tratada de um modo geral, como o significado de uma palavra, uma frase, uma sentença, ou qualquer texto em linguagem humana. Os sentidos mais estritos de semântica, como a web semântica ou abordagens semânticas baseadas na lógica formal não serão considerados. O foco está nos MEV e na sua relação com a *hipótese de distribuição* como uma abordagem para representar alguns aspectos da semântica da linguagem natural.

Criada em 1971 para o sistema de recuperação de informação SMART [41], a representação baseada em MEV tem como intenção representar cada documento em uma coleção como um ponto em um espaço (um vetor em um espaço vetorial). Pontos que estão próximos neste espaço são semanticamente similares e pontos que estão distantes

são semanticamente distantes. A pesquisa que se deseja fazer, na forma de uma consulta, é também representada como um ponto no mesmo espaço em que os documentos são inseridos (a consulta é considerada uma espécie de pseudo-documento). Os documentos são, então, classificados em ordem crescente de distância (ordem de diminuição da semelhança semântica) a partir da consulta.

O sucesso dos MEV em recuperação de informação tem inspirado pesquisadores a estender seu uso para outras tarefas de processamento de linguagem natural, com resultados expressivos: uso de uma representação vetorial para identificação do significado de palavras em questões envolvendo sinônimos, de múltipla escolha, no *Test of English as a Foreign Language (TOEFL)*, obtendo nota igual a 92,5% na prova cuja média humana é de 64,5% [39]; e uso de uma representação vetorial semelhante no exame de seleção do *SAT College* em questões de analogia, obtendo nota igual a 56%, próximo da média humana de 57% [51], por exemplo.

2.2 Motivação para o uso de MEV

A representação baseada em MEV apresenta muitas vantagens. O uso de MEV permite extrair automaticamente conhecimento de uma coleção de conteúdos, exigindo muito menos trabalho do que outras abordagens semânticas, tais como bases de conhecimento e ontologias codificadas manualmente.

Além disso, MEV apresenta excelente performance em tarefas que envolvem a medição de similaridade de significado entre palavras, frases e documentos. Algumas ferramentas de busca utilizam MEV para medir a similaridade entre uma consulta e uma coleção de documentos, indicando o documento que mais se aproxima da consulta apresentada [30]. Algoritmos para medir o relacionamento semântico [53] e a similaridade de relações semânticas [36] utilizam MEV.

Uma outra vantagem no uso de MEV é a relação que a reprentação possui com a *hipótese distributiva* e as hipóteses relacionadas. Na *hipótese distributiva*, palavras que ocorrem em contextos similares tendem a ter significados similares [11]. Esforços para aplicar essa hipótese em algoritmos concretos para medir a similaridade de significados muitas vezes levam a vetores, matrizes e tensores de ordem superior. Esta conexão entre a *hipótese distributiva* e MEV reforça a importância de se investigar o uso do modelo.

Apesar de o uso de vetores ser comum em ciência cognitiva e inteligência artificial, MEV introduziu a utilização de frequências em conjuntos de textos como pista para a descoberta de informações semânticas. Em aprendizado de máquina, um problema típico é aprender a classificar ou agrupar um conjunto de itens (ou seja, exemplos, casos,

indivíduos, entidades) representados como vetores de características [55]. Em geral, as características não são derivadas de frequências de eventos, embora isto seja possível [54].

Sistemas de recomendação também utilizam vetores [26]. Em um típico sistema de recomendação, temos uma matriz *pessoa-item* em que as linhas correspondem a pessoas (clientes, consumidores), as colunas correspondem aos itens (produtos, compras) e o valor de um elemento é a classificação (fraco, excelente) que a pessoa tenha dado ao item. Muitas das técnicas matemáticas que funcionam bem com matrizes *termo-documento* também funcionam bem com matrizes *pessoa-item*, mas as avaliações não são derivadas de frequências de eventos.

Na ciência cognitiva, a teoria de protótipo frequentemente faz uso de vetores. A idéia básica da teoria de protótipo é que alguns membros de uma categoria são mais centrais do que outros [22]. Por exemplo, o *pombo* é um membro central (protótipo) da categoria de *aves*, enquanto que *pinguim* é um membro mais periférico. Conceitos têm diferentes graus de adesão em categorias (categorização graduada). Uma maneira natural para formalizar tais observações é representar conceitos como vetores e categorias como conjuntos de vetores [48]. No entanto, estes vetores são geralmente baseados em contagens numéricas desencadeadas pelo questionamento a seres humanos; eles não são baseados em frequências de eventos.

Outra área da psicologia que faz uso extensivo de vetores é psicometria, que estuda a aferição de habilidades e traços psicológicos. O instrumento de aferição usual é um teste ou questionário, tais como testes de personalidade. Os resultados de um teste são tipicamente representados como uma matriz *indivíduo-item*, onde as linhas representam os indivíduos (pessoas) em um experimento e as colunas representam os itens (perguntas) do teste (questionário). O valor de um elemento na matriz é a resposta que o indivíduo forneceu para o item correspondente. Muitas técnicas de análise vetorial foram pioneiras em psicometria [54].

Na ciência cognitiva, Análise Semântica Latente (ASL) [24] está dentro do âmbito de MEV uma vez que utiliza modelos de espaço vetorial cujos valores dos elementos são derivados de frequências de eventos, tais como o número de vezes que uma determinada palavra aparece em um determinado contexto. Os cientistas cognitivos têm argumentado que há razões empíricas e teóricas para acreditar que MEV, como ASL, são modelos plausíveis sob alguns aspectos da cognição humana [23]. Em inteligência artificial, linguística computacional e recuperação de informação, essa plausibilidade não é essencial, mas pode ser vista como um sinal de que MEV é uma área promissora para futuras pesquisas.

2.3 Abordagem MEV

Na abordagem MEV discutida a seguir, $\mathbf{A}$ representa uma matriz, $\mathbf{a}$ representa um vetor e a representa um escalar. Essa notação base será utilizada para apresentar três representações MEV diferentes como forma de exemplificar o uso dessa abordagem na identificação de similaridade entre documentos, palavras e padrões: matriz *palavra-documento*, matriz *palavra-contexto* e matriz *par-padrão*.

Seja uma grande coleção de documentos e, consequentemente, um grande número de vetores-documento, torna-se conveniente a organização dos vetores em uma matriz. Os vetores-linha da matriz correspondem aos termos (geralmente termos são palavras, podendo existir outras possibilidades) e os vetores-coluna correspondem aos documentos (páginas web, por exemplo). Este tipo de matriz é chamada de matriz *termo-documento*.

Um multiconjunto (tradicionalmente conhecido como *bag*) corresponde a um conjunto onde duplicações são permitidas. Por exemplo, a, a, b, c, c, c é um multiconjunto contendo a, b, e c. A ordem dos elementos não importa em multiconjuntos; os multiconjuntos a, a, b, c, c, c e c, a, c, b, a, c são equivalentes. Podemos representar o multiconjunto a, a, b, c, c, c como um vetor $\mathbf{x} = \langle 2, 1, 3 \rangle$, considerando que o primeiro elemento de $\mathbf{x}$ é a frequência de a no multiconjunto, o segundo elemento é a frequência de b no multiconjunto e o terceiro elemento é a frequência de c no multiconjunto. Um conjunto de multiconjuntos pode ser representado como uma matriz $\mathbf{X}$, na qual cada coluna $\mathbf{x}_{:j}$ corresponde a um multiconjunto, cada linha $\mathbf{x}_{i:}$ corresponde a um único membro, e um elemento x_{ij} é a frequência do i-ésimo membro no j-ésimo multiconjunto.

Em uma matriz *termo-documento*, um vetor-documento representa o documento correspondente como um multiconjunto de palavras (*bag of words*). Em recuperação de informação, a *hipótese do multiconjunto de palavras* é que podemos estimar a relevância de documentos para uma determinada consulta através da representação dos documentos e da consulta como multiconjuntos de palavras. Em outras palavras, as frequências das palavras em um documento tendem a indicar a relevância do documento para uma consulta. A *hipótese do multiconjunto de palavras* é a base para a aplicação do MEV em recuperação de informação [43]. A hipótese expressa que um vetor-coluna em uma matriz *termo-documento* capta um aspecto do significado do documento correspondente; sobre o que é o documento.

Seja $\mathbf{X}$ uma matriz *termo-documento*. Suponha que a coleção contenha n documentos e m termos únicos. A matriz $\mathbf{X}$ terá, então, m linhas (cada linha representa um termo no vocabulário) e n colunas (cada coluna representa um documento). Suponha, ainda, que w_i represente o i-ésimo termo no vocabulário e que d_j represente o j-ésimo documento na coleção. A i-ésima linha em $\mathbf{X}$ é o vetor-linha $\mathbf{x}_{i:}$ e a j-ésima coluna em $\mathbf{X}$ é o vetor-coluna $\mathbf{x}_{:j}$. O vetor-linha $\mathbf{x}_{i:}$ contém n elementos, um elemento para cada documento, e

o vetor-coluna $\mathbf{x}_{:j}$ contém m elementos, um elemento para cada termo. Suponha que $\mathbf{X}$ é uma simples matriz de frequências. O elemento x_{ij} em $\mathbf{X}$ é a frequência do i-ésimo termo w_i no j-ésimo documento d_j.

Em geral, o valor da maior parte dos elementos de $\mathbf{X}$ será zero (matriz esparsa), uma vez que a maioria dos documentos utiliza apenas uma pequena fração de todo o vocabulário. Se forem escolhidos aleatoriamente um termo w_i e um documento d_j, é provável que w_i não ocorra em qualquer lugar de d_j e, portanto, x_{ij} é igual a 0.

O padrão de números em $\mathbf{x}_{i:}$ é uma espécie de *assinatura* do i-ésimo termo w_i. Do mesmo modo, o padrão de números em $\mathbf{x}_{:j}$ é uma *assinatura* do j-ésimo documento d_j. Em certo nível, o padrão de números diz sobre o que é determinado termo ou documento.

O vetor $\mathbf{x}_{:j}$ pode parecer uma representação bastante grosseira do documento d_j. Ela nos diz com que frequência as palavras aparecem no documento, mas a ordem sequencial das palavras é perdida. O vetor não captura a estrutura das frases, sentenças, parágrafos e capítulos do documento. No entanto, apesar dessa característica, os motores de busca funcionam surpreendentemente bem; vetores parecem capturar um aspecto importante da semântica [54].

Uma justificativa intuitiva para a matriz *termo-documento* é que o tema de um documento irá probabilisticamente influenciar a escolha de palavras pelo autor ao escrever o documento. Se dois documentos possuem temas semelhantes, então os dois vetores-coluna correspondentes tendem a ter padrões similares de valores.

A relevância de um documento para uma consulta é dada pela similaridade de seus vetores. É possível mudar o foco da análise para medir a similaridade entre palavras em vez da similaridade entre documentos, avaliando os vetores-linha da matriz *termo-documento* ao invés dos vetores-coluna [9]. Porém, um documento não é necessariamente o comprimento ótimo de texto para se medir a semelhança entre palavras. Em geral, é possível uma matriz *palavra-contexto* em que o contexto é dado por palavras, orações, frases, parágrafos, capítulos, documentos, ou possibilidades mais exóticas, como sequências de caracteres ou padrões.

Na *hipótese de distribuição*, em linguística, palavras que ocorrem em contextos semelhantes tendem a ter significados semelhantes [17]. Esta hipótese é a justificativa para a aplicação do MEV para medir a similaridade entre palavras. Uma palavra pode ser representada por um vetor em que os elementos são derivados das ocorrências da palavra em vários contextos, tal como janelas de palavras [28] e dependências gramaticais [37]. Vetores-linha semelhantes na matriz *palavra-contexto* indicam semelhança do significado das palavras.

Em uma matriz *par-padrão*, vetores-linha correspondem a pares de palavras como *pedreiro:pedra* e *carpinteiro:madeira*, e vetores-coluna correspondem aos padrões em que

ocorrem esses pares, tais como "X corta Y" e "X trabalha com Y". A matriz *par-padrão* foi introduzida com a finalidade de medir a similaridade semântica de padrões, isto é, a semelhança de vetores-coluna [25]. Dado um padrão tal como "X resolve Y", é possível encontrar padrões similares como "Y é resolvido por X", "Y é resolvido em X", e "X resolve Y".

A *hipótese de distribuição estendida* indica que os padrões que co-ocorrem com pares semelhantes tendem a ter significados semelhantes [25]. Os padrões "X resolve Y" e "Y é resolvido por X" tendem a co-ocorrer com pares semelhantes *X:Y*, sugerindo que tais padrões têm significados semelhantes. A similaridade de padrões pode ser usada para inferir que uma sentença é uma paráfrase de outra sentença.

A matriz *par-padrão* pode ser utilizada para medir a similaridade semântica das relações entre pares de palavras, isto é, a semelhança de vetores-linha [53]. Por exemplo, os pares de palavras *pedreiro:pedra*, *carpinteiro:madeira* e *vidraceiro:vidro* compartilham a relação semântica *artesão:material*. Em cada caso, o primeiro membro do par é um artesão que produz artefatos a partir do material indicado no segundo membro do par. Os pares tendem a co-ocorrer em padrões similares, como "o X é utilizado pelo Y para" e "o X transforma o Y em".

A *hipótese da relação latente* indica que pares de palavras que co-ocorrem em padrões semelhantes tendem a ter relações semânticas similares [52]. Pares de palavras com vetores-linha semelhantes em uma matriz *par-padrão* tendem a ter relações semânticas similares. Isto é o inverso da *hipótese de distribuição estendida* que indica que os padrões com vetores-coluna semelhantes na matriz *par-padrão* tendem a ter significados semelhantes.

As possibilidades não se esgotam com as matrizes *termo-documento*, *palavra-contexto* e *par-padrão*. Matrizes *trio-padrão* podem ser utilizadas para medir a similaridade semântica entre trios de palavras. Enquanto uma matriz *par-padrão* pode ter uma linha *pedreiro:pedra* e uma coluna "X trabalha com Y", uma matriz *trio-padrão* pode ter uma linha *pedreiro:pedra:alvenaria* e uma coluna "X usa Y para construir Z". No entanto, n-tuplas de palavras decrescem cada vez mais à medida que n aumenta. Por exemplo, frases que contenham as palavras *pedreiro*, *pedra* e *alvenaria*, juntas, são menos freqüentes do que frases que contenham as palavras *pedreiro* e *pedra* juntas. Uma matriz *trio-padrão* será muito mais esparsa do que uma matriz *par-padrão*. A quantidade de textos necessária a fim de se ter números suficientes para tornar as matrizes úteis cresce rapidamente à medida que n aumenta. Esse problema pode ser diminuído com a quebra de n-tuplas em pares. Por exemplo, *a:b:c* pode ser decomposto em *a:b*, *a:c* e *b:c*. A similaridade dos dois trios *a:b:c* e *d:e:f* pode ser estimada pela semelhança dos seus pares correspondentes. Uma matriz *par-padrão* relativamente densa poderia servir como um substituto para uma

matriz *trio-padrão* relativamente escassa.

2.4 Processamento Linguístico para MEV

A representação MEV requer, para que sejam obtidos melhores resultados, algumas etapas de pré-processamento linguístico para tratamento dos textos que serão representados de forma matricial. Os tipos de tratamentos que são utilizados podem ser agrupados em três classes. Em primeiro lugar, é recomendável *tokenizar* o texto bruto; ou seja, é preciso decidir o que constitui um termo e como extrair termos de texto bruto. Em segundo lugar, pode ser necessário *normalizar* o texto bruto para converter superficialmente diferentes cadeias de caracteres para uma mesma forma (por exemplo, *carro*, *Carro*, *Carros* e *carros* poderiam ser normalizados para *carro*). Por fim, pode ser interessante anotar o texto bruto para marcar strings idênticos como sendo diferentes (por exemplo, *lista* como um verbo pode ser anotado como *lista/VB* e *lista* como um substantivo pode ser anotado como *lista/NN*).

2.4.1 Tokenização

A tokenização consiste em identificar cada palavra ou termo que compõe um texto, envolvendo o tratamento de caracteres especiais, tais como espaços em branco e pontuação (hífen, apóstrofo, etc), e a remoção das chamadas *stop words* funcionais, palavras que ocorrem com alta frequência mas que agregam pouco significado ao contexto, tais como alguns pronomes e artigos.

Em algumas línguas (Chinês, por exemplo), as palavras não são separadas por espaços. Nestes casos, a representação MEV necessita quebrar o texto em unigramas ou bigramas.

2.4.2 Normalização

A motivação para a normalização é a observação de que muitas cadeias diferentes de caracteres transmitem significados semelhantes. Uma vez que o objetivo é chegar ao significado que está "por trás das palavras", parece razoável normalizar variações superficiais de um mesmo termo. Os tipos mais comuns de normalização são representar todos os caracteres em letra maiúscula e reduzir palavras flexionadas à sua forma raiz.

Representar caracteres em letra maiúscula pode ser problemático em alguns idiomas. No Francês, por exemplo, acentos são opcionais para letras maiúscula, podendo ser difícil recuperar acentos ao converter palavras em minúsculas. Algumas palavras não podem ser distinguidas sem acentos; por exemplo, *PECHE* poderia ser tanto *pêche* (pêssego) ou *péché* (pecado).

A morfologia é o estudo da estrutura interna das palavras. Muitas vezes uma palavra é constituída por uma raiz com afixos adicionados (inflexões), tais como as formas plural e tempos no passado. Uma espécie de análise morfológica é o processo de reduzir palavras flexionadas para sua raiz. Em Inglês, afixos são mais simples e mais regulares do que em muitas outras línguas, e algoritmos baseados em heurísticas funcionam relativamente bem.

Em geral, a normalização provoca um aumento no *recall* e uma diminuição na *precisão*. Quando removemos variações superficiais que aparentemente são irrelevantes para o significado, torna-se mais fácil reconhecer similaridades. Mas às vezes essas variações superficiais têm significado semântico relevante, diminuindo a *precisão*.

2.4.3 Anotação

Anotação é o inverso da normalização. Assim como diferentes sequências de caracteres podem ter o mesmo significado, cadeias de caracteres idênticas podem ter significados diferentes, dependendo do contexto. As formas mais comuns de anotação incluem análise *part-of-speech* (marcação das palavras de acordo com as suas partes do discurso), sentido da palavra (marcação das palavras ambíguas de acordo com seus significados) e análise gramatical (análise da estrutura gramatical das sentenças).

Uma vez que a anotação é o inverso da normalização, espera-se uma diminuição do *recall* e um aumento da *precisão*. Por exemplo, marcar *programa* como um substantivo ou verbo possibilita pesquisar de forma seletiva os documentos que tratam do ato de *programação de computadores* (verbo) em vez de documentos que abordam determinados *programas de computador* (substantivo), aumentando a *precisão*. No entanto, um documento sobre *programas de computador* (substantivo) pode ter algo de útil a dizer sobre o ato de *programação de computadores* (verbo), mesmo que o documento nunca use a forma verbal do *programa*, diminuindo o *recall*.

2.5 Processamento Matemático para MEV

Após a tokenização e (opcionalmente) normalização e anotação, o primeiro passo é a geração de uma matriz de frequências. Em segundo lugar pode ser necessário ajustar os pesos dos elementos da matriz, porque as palavras comuns terão alta freqüência, no entanto, são menos informativas do que as palavras raras. Em terceiro lugar pode ser necessário suavizar a matriz para reduzir a quantidade de ruídos e preencher alguns elementos contendo zero numa matriz esparsa. Por fim, há muitas maneiras diferentes de se medir a semelhança entre dois vetores.

Como visto acima, a construção de um MEV pode ser descrita como sendo um processo de quatro etapas: calcular as frequências, aplicar pesos, suavizar (redução de dimensionalidade) e calcular as semelhanças [27].

2.5.1 Construção da Matriz de Frequências

Um elemento em uma matriz de freqüências corresponde a um evento: um determinado item (termo, palavra, par) que ocorre em uma determinada situação (documento, contexto, padrão) um certo número de vezes (frequência). Em um nível abstrato, a construção de uma matriz de frequências é uma simples questão de contagem de eventos. Na prática, isto pode ser complicado quando o conjunto de documentos é grande.

Uma abordagem típica para a construção de uma matriz de frequências envolve duas etapas. Em primeiro lugar verifica-se sequencialmente o conjunto de documentos, armazenando os eventos e suas frequências em uma tabela, um banco de dados ou um índice motor de busca. Em segundo lugar, usa-se a estrutura de dados resultante para gerar a matriz de frequências, com uma representação de matriz esparsa [13].

2.5.2 Ponderação

A ideia da ponderação é atribuir maior peso aos eventos surpreendentes e menos peso aos eventos esperados. A hipótese é de que os eventos surpreendentes, se compartilhados por dois vetores, são mais discriminativos da semelhança entre os vetores do que eventos menos surpreendentes. Por exemplo, na medição da semelhança semântica entre as palavras *rato* e *camundongo*, os contextos *dissecar* e *exterminar* são mais discriminativos de sua semelhança do que os contextos *têm* e *gostam*. Em teoria da informação, um evento surpreendente tem mais conteúdo que um evento esperado [45].

A forma mais popular de formalizar esta ideia para matrizes *termo-documento* é a família TF-IDF (frequência do termo *versus* inverso da frequência no documento) de funções de ponderação [19]. Um elemento recebe um peso elevado quando o termo correspondente é frequente no documento correspondente (ou seja, TF é alta), mas o termo é raro em outros documentos do conjunto (ou seja, DF é baixa e, assim, IDF é alta). As funções de ponderação da família TF-IDF podem produzir melhoras significativas em tarefas de recuperação de informação quando comparadas com a frequência bruta [42].

A $tf_{t,d}$ representa a frequência do termo t no documento d. A df_t de um termo é uma medida inversa da informatividade do termo t. Representa o número de documentos em que um termo aparece, isto é, df_t não pode ser maior do que o número de documentos (N).

O inverso da frequência do documento pode ser calculado da seguinte forma:

$$idf_t = \log_{10} \frac{N}{df_t} \tag{2.1}$$

A função *log* é utilizada em vez de N/df_t para "amortecer" o efeito da idf_t. Por exemplo, se temos 1 milhão de documentos e um termo que aparece apenas em um documento, então idf_t é igual a 6, ou seja, temos um termo relevante. Por outro lado, se tivermos um termo que aparece em todos os documentos, a idf_t será 0, indicando tratar-se de um termo comum que provavelmente não é relevante.

Podemos aplicar este mesmo conceito para obter o termo ponderado de *tf*, neste caso, o coeficiente da frequência logarítmica ($w_{t,d}$) de um termo t em um documento d, que é definido por:

$$w_{t,d} = 1 + \log_{10} tf_{t,d} \tag{2.2}$$

Finalmente, a TF-IDF ponderada de um termo é o produto das componentes *tf* ponderada e *idf* ponderada:

$$w_{t,d} = (1 + \log_{10} tf_{t,d}) \times log_{10} \frac{N}{df_t} \tag{2.3}$$

Outro tipo de ponderação, muitas vezes combinada com a ponderação TF-IDF, é a normalização de comprimento. Em recuperação de informação, se o comprimento do documento for ignorado, os motores de busca tendem a ter um viés em favor de documentos mais longos. A normalização do comprimento corrige esse viés [47].

A ponderação de termos também pode ser utilizada para corrigir termos correlacionados. Por exemplo, os termos *refém* e *reféns* tendem a ser correlacionados, ainda assim, pode não ser recomendável normalizá-los para o mesmo termo porque eles possuem significados ligeiramente diferentes. Como uma alternativa para a normalização, pode-se reduzir seus pesos quando quando ambos ocorrerem em um documento [6].

2.5.3 Suavização

A maneira mais simples de melhorar o desempenho de um processo de recuperação de informação é limitar o número de componentes do vetor. Manter apenas as partes que representem palavras de conteúdo que ocorrem mais frequentemente é uma maneira; no entanto, as palavras comuns, tais como *o* e *têm*, possuem pouco poder de discriminação semântica. Componentes simples de suavização heurística, com base nas propriedades de esquemas de ponderação, têm demonstrado tanto a manutenção do poder de discriminação semântica quanto a melhora do desempenho de cálculos de similaridade.

Calcular a semelhança entre todos os pares de vetores é uma tarefa computacionalmente intensiva. No entanto, apenas os vetores que compartilham uma coordenada não-zero devem ser comparados (isto é, dois vetores que não compartilham nenhuma coordenada são diferentes).

2.5.4 Cálculo de Similaridade

A maneira mais popular de se medir a similaridade entre dois vetores é através do cálculo do cosseno entre eles. Sejam $\mathbf{x}$ e $\mathbf{y}$ dois vetores, cada um com n elementos:

$$\begin{aligned} \mathbf{x} &= \langle x_1, x_2, ..., x_n \rangle \\ \mathbf{y} &= \langle y_1, y_2, ..., y_n \rangle \end{aligned} \tag{2.4}$$

O cosseno do ângulo entre $\mathbf{x}$ e $\mathbf{y}$ pode ser calculado da seguinte forma:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}} \tag{2.5}$$

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\sqrt{\mathbf{x} \cdot \mathbf{y}} \cdot \sqrt{\mathbf{x} \cdot \mathbf{y}}} \tag{2.6}$$

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}}{\|\mathbf{x}\|} \cdot \frac{\mathbf{y}}{\|\mathbf{y}\|} \tag{2.7}$$

Em outras palavras, o cosseno do ângulo entre dois vetores é o produto interno desses vetores depois da normalização (vetores unitários). Se $\mathbf{x}$ e $\mathbf{y}$ são vetores de frequências de palavras, palavras comuns terão longos vetores enquanto que palavras raras terão vetores curtos, mesmo se forem sinônimos. O cosseno capta a ideia de que o comprimento dos vetores é irrelevante; o importante é o ângulo entre eles.

O cosseno varia entre -1 quando os vetores apontam em sentidos opostos (o ângulo é de 180 graus) e 1 quando apontam na mesma direção (o ângulo é 0). Quando os vetores são ortogonais (o ângulo é de 90 graus), o cosseno é zero. Normalmente quando não existem elementos negativos nos vetores, o cosseno também não é negativo, mas a ponderação e a suavização podem introduzir elementos negativos.

Uma medida da distância entre os vetores pode ser facilmente convertida em uma medida de similaridade por inversão ou subtração:

$$sim(\mathbf{x}, \mathbf{y}) = 1/dist(\mathbf{x}, \mathbf{y}) \tag{2.8}$$

$$sim(\mathbf{x}, \mathbf{y}) = 1 - dist(\mathbf{x}, \mathbf{y}) \tag{2.9}$$

Muitas medidas de similaridade foram propostas em recuperação de informação [20]. Costuma-se dizer que, com vetores devidamente normalizados, a diferença de desempenho da recuperação usando diferentes medidas é insignificante [40].

2.6 Principais Aplicações de MEV

Existem diversas aplicações para a representação MEV. Nesse trabalho, MEV será utilizado para a realização de ligação de entidades de nomes próprios e conceitos concretos, porém essa representação pode ser empregada também em:

- Recuperação de documentos: a ideia central é, dada uma consulta, estabelecer um ranking decrescente conforme o cosseno entre os vetores-documento e o vetor-consulta. Outra utilização é em recuperação de documentos em diferentes idiomas, onde uma consulta em um idioma é utilizada em um vetor de documentos em outro idioma.
- Agrupamento de documentos: dada uma medida de similaridade de documentos, os documentos são agrupados de forma que em um determinado grupo a similaridade entre documentos tende a ser alta, enquanto que entre os grupos a similaridade tende a ser baixa.
- Classificação de documentos: dado um conjunto de documentos rotulados para treinamento, a tarefa é aprender como rotular documentos de um conjunto de teste. Os rótulos podem ser tópicos de um documento, sentimento dos documentos, etc.
- Avaliação de ensaios: essa tarefa pode ser realizada calculando-se a similaridade de um ensaio com outros ensaios de referência, representados em uma matriz *palavra-documento*.
- Segmentação de documentos: dividir um documento em subtópicos. Os documentos são tratados como um conjunto de blocos em uma representação de matriz *palavra-bloco*.

2.7 Métricas de Avaliação

Nesta seção, são apresentadas duas métricas recorrentemente utilizadas na análise de desempenho de algoritmos para a resolução de problemas envolvendo recuperação de informação e ligação de entidades: *precisão* e *recall*.

2.7.1 *Precisão*

No campo da recuperação de informação, *precisão* é a fração de documentos recuperados que são relevantes [57]. *Precisão* também é chamada de valor preditivo positivo. Essa métrica pode ser calculada conforme segue:

$$Precisão = \frac{\|documentos - relevantes\| \cap \|documentos - obtidos\|}{\|documentos - obtidos\|} \tag{2.10}$$

A *precisão* pode ser compreendida como o número de verdadeiros positivos (ou seja, o número de itens corretamente rotulados como pertencentes à classe positiva) dividido pelo número total de elementos rotulados como positivos (ou seja, a soma dos verdadeiros positivos e falsos positivos). Por exemplo, para o resultado de uma busca textual em um conjunto de documentos, *precisão* é o número de resultados corretos divididos pelo número de todos os resultados retornados.

Considerando essa definição, pode-se constatar que maximizar a *precisão* significa minimizar os falsos positivos.

2.7.2 *Recall*

No campo da recuperação de informação, *recall* é a fração de documentos relevantes que são recuperados [57]. *Recall* também é chamado de sensibilidade. Essa métrica pode ser calculada conforme segue:

$$Recall = \frac{\|documentos - relevantes\| \cap \|documentos - obtidos\|}{\|documentos - relevantes\|} \tag{2.11}$$

O *recall* pode ser compreendido como o número de verdadeiros positivos dividido pelo número total de elementos que efetivamente pertencem à classe positiva (ou seja, a soma de verdadeiros positivos e falsos negativos). Por exemplo, para o resultado de uma busca textual em um conjunto de documentos, *recall* é o número de resultados corretos dividido pelo número de resultados que deveriam ter sido retornados.

Considerando essa definição, pode-se constatar que maximizar o *recall* significa minimizar os falsos negativos.

2.8 Resumo do Capítulo

O presente capítulo apresentou a fundamentação teórica citando os principais conceitos relacionados ao Modelo de Espaço Vetorial, utilizado no desenvolvimento deste trabalho. Nesta parte foram abordados brevemente: conceitos iniciais, motivação, fundamentação teórica/matemática e principais problemas abordados com o uso de MEV.

A aplicabilidade de MEV é verificada na representação e seleção de entidades que apresentam maior similaridade com os nomes próprios e conceitos concretos extraídos de textos. Os conceitos que envolvem a metodologia, tais como tokenização, normalização e anotação, são aplicados neste trabalho como forma de tratar os textos de onde as menções são extraídas, com o objetivo de formar um grupo homogêneo de entrada, aumentando a chance de acertos.

O processamento matemático apresentado neste capítulo é retomado no capítulo que trata da descrição do modelo, onde são fornecidas as definições empregadas no caso específico do cálculo da similaridade semântica entre entidades e menções (nomes próprios e conceitos concretos). Os conceitos apresentados neste capítulo são aplicados no presente trabalho quase que em sua totalidade, com pequenos ajustes na representação explicados em detalhes nos capítulos seguintes.

Capítulo 3

Ligação de Entidades

Neste capítulo são apresentados os fundamentos relacionados à ligação de entidades. Na seção 3.1 são apresentados os conceitos iniciais, enquanto que na seção 3.2 são descritos os principais problemas abordados pela área. A seção 3.3 trata da importância das bases de entidades para a teoria. A seção 3.4 dedica-se ao tratamento das principais classes de entidades investigadas. O capítulo conclui com os desafios relacionados à ligação de entidades, na seção 3.5, e com a relação entre ligação de entidades e conceitos concretos, na seção 3.6. Por fim, a seção 3.7 traz um resumo do capítulo.

3.1 Conceitos Iniciais

Referências a entidades como pessoas, lugares e organizações nos textos são difíceis de se detectar de maneira automática, porque as entidades podem fazer referências a diferentes *strings*, e uma mesma *string* pode fazer referência a várias entidades. Por exemplo, *David Murray* pode referir-se tanto ao saxofonista de jazz ou o guitarrista do Iron Maiden, que pode ser conhecido por outros apelidos como *Mad Murray*. Estes problemas sinonímia e ambiguidade tornam a coleta e a exploração de informações sobre entidades em documentos difícil por parte dos sistemas de processamento de linguagem, sem primeiro ocorrer uma ligação dessas entidades com uma base de dados [15].

Ligação de Entidades (LE) é a tarefa de relacionar entidades mencionadas com as respectivas entradas em bases de conhecimento. LE é útil quando se deseja referenciar diretamente pessoas, lugares e organizações, sem se preocupar com caracteres ambíguos ou redundantes. No domínio das finanças, LE pode ser usado para conectar informações textuais sobre as empresas a dados financeiros, por exemplo, notícias e mercados [35]. LE também pode ser usada em pesquisa, onde os resultados de consultas por entidades podem incluir dados tradicionais sobre a entidade, além de páginas que falam sobre o mesmo assunto [3].

As principais abordagens para a realização de LE, mesmo que de forma implícita, fazem o uso de três importantes passos no processo de tratamento dos documentos e seleção dos resultados. Os passos são:

1. Detectar as menções (sentenças "ligáveis");
2. Classificar e selecionar as respectivas entidades; e
3. Desambiguar/melhorar os resultados com base no contexto (semântica).

Apesar da existência de um roteiro semelhante para a realização de ligação de entidades, as pesquisas mais recentes apresentam diferenças entre si em cada um dos itens citados anteriormente. Existem diferentes abordagens, tanto para detecção de menções como para a desambiguação e seleção de entidades.

LE é semelhante ao problema amplamente estudado de desambiguação de sentidos (*word sense disambiguation - WSD)* com a Wikipédia no papel da WordNet [14]. Em essência, ambos resolvem problemas de sinonímia e ambiguidade na linguagem natural. As tarefas diferem em termos de pesquisa de candidatas e detecção de entidades nulas. WSD assume que a WordNet é um recurso completo e consistente para encontrar possíveis referências para uma dada palavra. A mesma abordagem é aplicada na *wikificação* onde frases arbitrárias contendo nomes e termos mais gerais são comparadas com páginas da Wikipédia [21]. No entanto, não é fornecido um mecanismo para tratar objetos que não estão presentes na base de dados. LE, por outro lado, não assume que a base de dados é completa, exigindo o tratamento de entidades mesmo não existentes na base [31]. Além disso, as entidades identificadas podem apresentar uma maior variedade de formas do que as tratadas em WSD, dificultando o processo de LE.

Até recentemente não era possível alcançar domínios específicos devido ausência de informações publicamente disponíveis sobre entidades. No entanto, a Wikipédia tem emergido como um importante repositório de conhecimento coletivo semi-estruturado sobre entidades diversas. Por conseguinte, tem sido amplamente utilizado para a modelagem do conhecimento [38].

Os conjuntos de dados mais populares para LE foram distribuídos por ocasião do *Knownledge Base Population* na NIST *Text Analysis Conference (TAC)*. Os participantes, em 2009, desenvolveram sistemas que ligavam um conjunto de 3.904 entidades mencionadas em notícias e textos da web com uma base formada de infoboxes da Wikipédia. A popularização dos estudos em LE alcançou diversas aplicações e conjuntos de dados, que serão comentados na próxima seção.

3.2 Aplicações envolvendo Ligação de Entidades

Várias comunidades de pesquisa têm discutido o problema de ambiguidade de entidades nomeadas, tendo sido abordado de duas maneiras distintas. No ramo de linguística computacional, o problema foi concebido pela primeira vez como uma extensão do problema de resolução de correferência [2]. Posteriormente a Wikipédia foi introduzida para auxiliar a desambiguação por meio das ligações entre as páginas que, em muitas das vezes, lidam com links ambíguos [33]. Finalmente, o pré-processamento foi aliado ao uso da Wikipédia para a obtenção de ligações para todas as entidades nomeadas, mesmo que nulas [3].

3.2.1 Geração de Dados com Pseudo-Palavras

Devido à dificuldade na anotação manual de dados, a estratégia de se adotar pseudo-palavras para gerar desambiguação artificial de sentidos tem despertado o interesse [12]. O dado é gerado a partir de duas palavras que não possuem sentido ambíguo, e todas as instâncias são substituídas por chaves ambíguas. Por exemplo, todas as instâncias da palavra *banana* e *porta* são substituídas pela chave ambígua *banana-porta*. A versão original, inequívoca, é reservada como um padrão para treinamento e avaliação.

Os dados para a resolução de correferência entre documentos podem ser gerados da mesma maneira, tomando todas as instâncias de duas ou mais palavras e misturando-as sob uma chave de anonimização como *Pessoa X*. A tarefa é, então, agrupar os documentos de acordo com as palavras originais [29].

A geração de pseudo-palavras é problemática tanto para desambiguação de sentidos como para desambiguação de entidades, mas por razões diferentes. Na desambiguação de sentidos as maiores ambiguidades ocorrem entre significados relacionados. Por exemplo, os significados de *tênis* e *matemática* da palavra *set* podem ser ligados de volta para um conceito compartilhado. Poucas ambiguidades de sentidos ocorrem entre conceitos não relacionados, tais como *banana* e *porta*, e é muito difícil selecionar pares de palavras que refletem as relações significativas entre os sentidos.

Na desambiguação de entidades há pouca razão para acreditar que duas pessoas chamadas *John Smith* irão partilhar mais propriedades do que uma entidade *Paul Simonell* e outra *Hugh Diamoni*, de maneira que a crítica feita às pseudo-palavras para desambiguação de sentidos não ocorre. Por outro lado, as entidades têm estruturas internas interessantes que um sistema de desambiguação pode explorar. Por exemplo, a utilização de um título como *Senhor* e *Doutor* pode ser relevante na caracterização de entidades complexas.

3.2.2 Wikificação

O desenvolvimento da Wikipédia ofereceu uma nova maneira de abordar o problema da desambiguação de entidades. Wikificação consiste em adicionar links a partir de conceitos importantes mencionados em textos para artigos da Wikipédia correspondentes. A tarefa difere da LE na ligação de conceitos que não são, necessariamente, entidades, e na base de conhecimento que é considerada completa.

3.2.3 Ligação de Entidades Nomeadas

A ligação de entidades nomeadas dedica-se à localização e classificação de palavras e/ou sentenças extraídas de um texto em categorias pré-definidas, tais como nomes de pessoas, organizações, lugares, etc. As primeiras tentativas de ligação de entidades nomeadas (LEN) - a tarefa de ligar entidades mencionadas em textos a uma base de dados - tinham como objetivo os links da Wikipédia. Estudos demonstraram que a ambiguidade dos links da Wikipédia é muito menor do que a ambiguidade de entidades mencionadas em textos de notícias [8]. Uma das possíveis causas dessa característica é que, para facilitar a recuperação de arquivos, os editores são encorajados a escolher uma terminologia mais consistente para anconrar o texto.

3.2.4 Estudos Recentes

A tarefa de ligação de entidades está diretamente relacionada a outros desafios em teoria da informação, dentre os quais podemos destacar: classificação de documentos; e identificação de entidades nomeadas.

De maneira mais simplificada, o problema de ligação de entidades consiste na ligação de sentenças/palavras extraídas de diferentes tipos de texto (documentos, blogs, tweets, etc) às suas respectivas entidades, entidades estas tipicamente extraídas de bases de conhecimento tais como Wikipédia e Freebase, por exemplo.

Dentre as principais aplicações abordadas pelos estudos mais recentes em LE destacam-se:

- pesquisa semântica;
- experiência do usuário (interface);
- melhoria automática de documentos;
- leitura direta (*go-read-here*);
- anotações *inline* (RDA, por exemplo);

Figura 3.1: Exemplo de ligação de entidades (TagMe). Fonte: http://tagme.di.unipi.it, acessado em 20/06/2015, às 15h40.

- aprendizado de ontologias;
- população de bases de conhecimento;
- redução dimensional (vetores de termos); e
- melhoria na classificação, recuperação, desambiguação e similaridade semântica de documentos, dentre outras.

Um exemplo desses estudos recentes pode ser visualizado na figura 3.1. Trata-se de um exemplo de resposta da ferramenta TagMe que recebeu, neste caso, como entrada, um texto sobre o presidente Barack Obama e, de maneira automática, sugeriu o link cujo conteúdo encontra-se destacado no quadro azul claro como entidade para a menção "Barack Obama" que, para ferramenta, mereceu o destaque e a ligação ao link sugerido.

3.2.5 Classes de problemas

Os principais problemas relacionados à extração de informação e à anotação de entidades, extraídas da Wikipédia, podem ser categorizados em três classes principais:

- *Disambiguate to Wikipedia* (D2W): consiste na escolha da entidade mais adequada para cada menção;
- *Scored-annotate to Wikipedia* (Sa2W): semelhante ao problema A2W, porém leva em consideração a atribuição de notas às entidades candidatas;

É possível estabelecer a seguinte relação entre os problemas listados acima e os anotadores de entidades mais recentes: AIDA (Sa2W, D2W); Illinois Wikifier (Sa2W, D2W); DBpedia Spotlight (Sa2W); TagMe (Sa2W); e Wikipedia Miner (Sa2W) [7]. Esses anotadores serão vistos com mais detalhes no capítulo 4. Na tabela 3.1 são apresentados alguns exemplos para as categorias de problemas em questão:

Tabela 3.1: Exemplos de problemas envolvendo anotação de entidades.

Problema	**Entrada**	**Saída**
D2W	A **história** começa no Condado, onde o Hobbit Frodo Baggins recebe **o** Anel de **Bilbo**.	História *null* Bilbo Baggins
Sa2W	A **história** começa no **Condado**, onde o **Hobbit Frodo Baggins** recebe **o Anel** de **Bilbo**.	História (0,8) Condado (Terra-Média) (0,5) Hobbit (1,0) Bilbo Baggins (1,0) O Anel (0,5) Bilbo Baggins (0,7)

A tabela 3.1 apresenta exemplos para os probelmas D2W e Sa2W. No exemplo *disambiguate to Wikipedia* (D2W) para um dado texto de entrada são identificadas as principais menções encontradas no texto e as respectivas entidades, que inclusive pode ter como resultado a atribuição de *null*. Já para o exemplo *scored-annotate to Wikipedia* (Sa2W) também são identificadas as principais menções e as respectivas entidades, com a diferença de que cada entidade recebe uma nota que representa a relevância da entidade para a menção (número indicado entre parênteses).

3.3 Bases de Entidades

Os frameworks para a realização de ligação de entidades, na maioria das vezes, fazem uso de bases de entidades tanto na atividade de desambiguação de entidades como no fornecimento de entidades candidatas. Dessa forma, a escolha de determinada base de conhecimento é fundamental para a obtenção de resultados satisfatórios. Atualmente, as bases de entidades mais utilizadas pelos pesquisadores são a Wikipédia, a DBpedia, a Freebase e a YAGO.

A principal base de conhecimento para a ligação de entidades é, sem sombra de dúvidas, a Wikipédia. Sua importância é tão grande que as classes de problemas citadas anteriormente foram definidas justamente em termos do uso da Wikipédia. A vantagem da Wikipédia reside na grande quantidade de conteúdo disponível, sobre os mais variados assuntos, atualizado diariamente, que pode ser acessado de maneira relativamente simples uma vez que os artigos guardam certa estrutura e estão disponíveis em sua totalidade também na versão *offline*.

Além dos artigos, a Wikipédia oferece: páginas de redirecionamentos, que estabelecem conexões entre artigos; páginas de desambiguação, que apresentam diferentes opções de artigos para uma mesma consulta; categorias de artigos, que auxiliam na classificação de

conteúdos; e hiperlinks, que estabelecem conexões entre sentenças de determinado artigo com outras páginas da Wikipédia. A figura 3.2 apresenta um artigo retirado da Wikipédia justamente abordando o tema base de conhecimento. Porém, o objetivo da figura é ilustrar como as informações contidas na Wikipédia sobre os mais variados assuntos podem ser úteis para os estudos em LE.

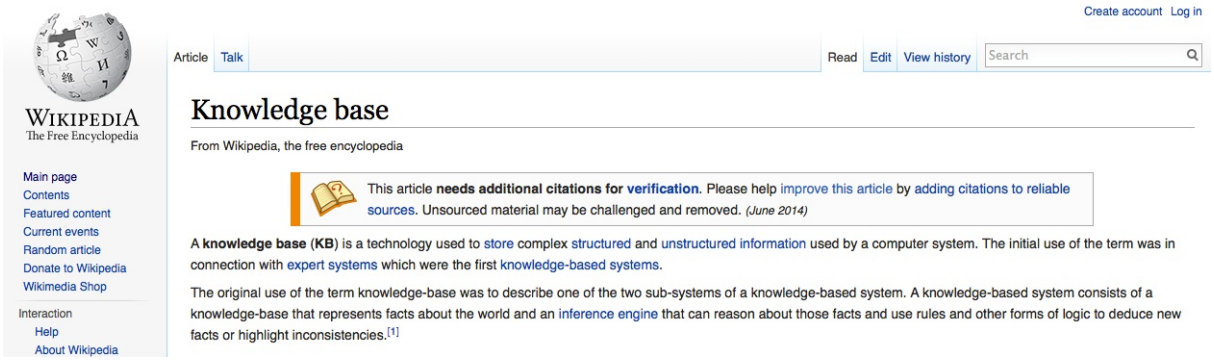

Figura 3.2: Wikipédia como base de conhecimento. Fonte: http://en.wikipedia.org/wiki/Knowledge_base, acessado em 20/06/2015, às 17h50.

As páginas da Wikipédia possuem uma série de atributos, além dos citados anteriormente, que podem ser explorados visando melhorar o processo de desambiguação. Dentre esses atributos podemos destacar:

- *Títulos* - o título canônico dos artigos.
- *Títulos de redirecionamento* - a Wikipédia fornece um mecanismo de redirecionamento automático a partir de títulos não canônicos - tais como variações de escrita, abreviações, línguas estrangeiras, etc - para artigos relevantes.
- *Termos do primeiro parágrafo em destaque* - palavras comuns para determinado tópico são convencionalmente inseridas em destaque no primeiro parágrafo.
- *Textos âncoras* - links entre páginas da Wikipédia. Textos âncora oferecem uma variedade de formas usadas como referências para menções do texto em que se encontram.
- *Títulos de páginas de desambiguação* - lista dos artigos que podem corresponder a um título ambíguo. Páginas de desambiguação normalmente consistem em uma ou mais listas, cada lista com itens ligados às páginas candidatas referentes ao termo ambíguo pesquisado.

- *Textos destacados de redirecionamento* - uma página pode desambiguar múltiplos termos - por exemplo, podemos ter uma página de desambiguação para *AMP* e outra para *Amp*.
- *Títulos truncados* - a Wikipédia convencionalmente acrescenta frases de desambiguação para formar um título único, tal como em "John Howard (Australian actor)".

Por sua vez, a DBpedia é produto de um esforço da comunidade para extrair informações mais estruturadas da Wikipédia e tornar essa informação disponível na web. Dessa forma, é possível realizar consultas mais sofisticadas ao conteúdo disponível na enciclopédia eletrônica e estabelecer conexões de diferentes conjuntos de dados na web com a Wikipédia. O principal objetivo do projeto DBpedia é tornar mais fácil a utilização dessa grande quantidade de informação, inspirando novas pesquisas na área de inteligência da web.

A Freebase, um projeto adquirido pelo Google em 2010, consiste em uma base de conhecimento armazenada em forma de grafo contendo 10 milhões de tópicos, milhares de tipos e dezenas de milhares de propriedades relacionadas a diversos conteúdos. Devido a sua estrutura, é possível estabelecer diversas conexões entre diferentes tipos de informações, possibilitando diferentes perspectivas ("Bob Dylan" foi um compositor, cantor, escritor e ator). Todo esse conteúdo está disponível na web e pode ser utilizado a partir da API da própria Freebase. Na figura 3.3 é exibido um exemplo da categorização que a Freebase possui dos dados armazenados em seu repositório.

Já a YAGO consiste em uma grande base de conhecimento semântico, derivada da Wikipédia, WordNet e GeoNames. Atualmente, a base contém mais de 10 milhões de entidades e mais de 120 milhões de fatos relacionados a essas entidades. Lançando mão da taxonomia da WordNet, YAGO permite a classificação das entidades em mais de 350 mil classes diferentes, possuindo dimensões temporal e espacial para muitos de seus fatos e entidades.

3.4 Classes de entidades

Os primeiros trabalhos em teoria da informação dedicavam-se principalmente a tratar conceitos mais bem sintatica e semanticamente definidos, tais como nomes de pessoas, lugares, organizações, datas, etc. O aprofundamento nesse universo provocou o surgimento de ferramentas com excelentes desempenhos quando avaliada a relevância dos resultados.

Por outro lado, o aumento da quantidade de conteúdo disponível na web fez surgir a necessidade de expandir a quantidade de classes de entidades a serem manipuladas pelos frameworks de tratamento de informação. Seguindo essa tendência, as bases de

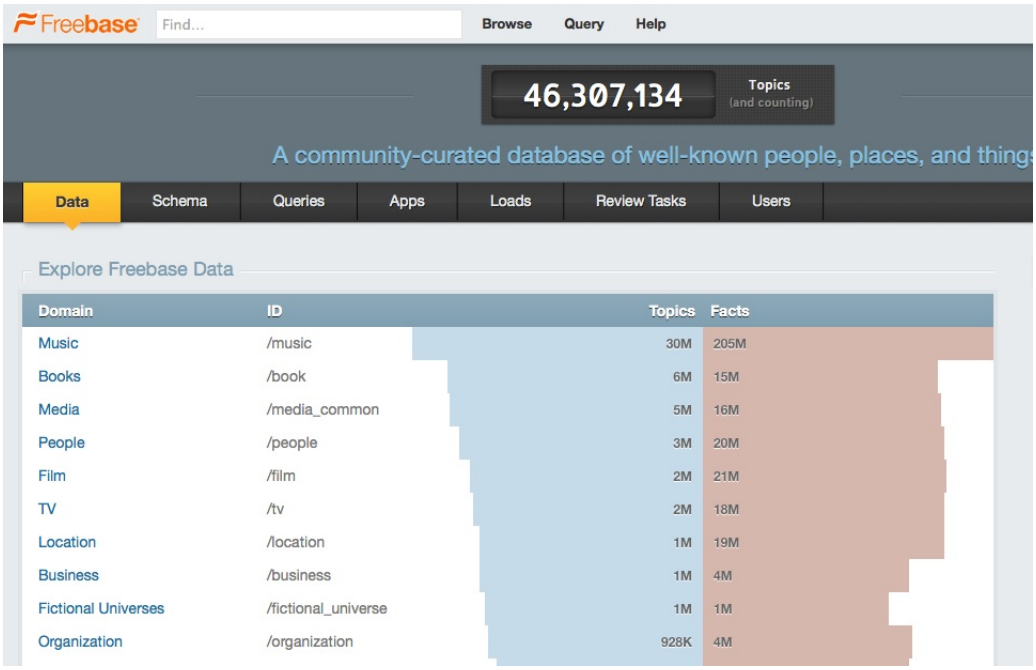

Figura 3.3: Freebase: milhares de tipos de entidades. Fonte: https://www.freebase.com, acessado em 21/06/2015, às 11h20.

conhecimento começaram a evoluir em dois sentidos principais: fortalecer a estrutura de armazenamento das informações; e aumentar as classes de entidades disponíveis nas bases de conhecimento. Um exemplo dessa evolução foi citado na seção anterior, a YAGO, que conta com centenas de milhares de classes de informações diferentes. Essa evolução, tanto no domínio dos problemas como nas abordagens utilizadas serão melhor demonstradas no capítulo 4.

3.5 Desafios relacionados à ligação de entidades

Apesar dos avanços obtidos pelas pesquisas mais recentes na área de ligação de entidades, existem ainda alguns assuntos em aberto cujas pesquisas relacionadas tem muito a evoluir. Dentre eles podemos citar:

- ligação de entidades multi-idiomas/entre idiomas;
- ligação de entidades entre bases de dados;
- ligação de entidades envolvendo coleções genéricas de textos/testes;
- ligação de entidades "além das entidades"; e
- incorporação de evidências contextuais no processo de ligação de entidades.

O presente estudo, como será melhor detalhado na próxima seção e nos capítulos seguintes, dedica-se ao processo de ligação de conceitos concretos, ou seja, sentenças que podem não se enquadrar nas categorias tradicionais de classes de entidades. Dessa forma, da lista acima, o desafio mais aderente ao presente estudo é aquele citado no item "ligação de entidades além das entidades".

3.6 Ligação de entidades e conceitos concretos

A definição sugerida para conceitos concretos indica que as entidades mencionadas nos textos podem não pertencer a nenhum tipo de classes de entidades específicas, por mais que existam centenas de classes de entidades. Isso ocorre porque, ao restringirmos o universo de busca por entidades mencionadas a um grupo específico de classes de entidades, podemos estar sujeito à perda de informações.

No presente trabalho as entidades nomeadas serão identificadas como pertencentes a somente dois grandes grupos de classes de entidades: nomes próprios e conceitos concretos. As entidades nomeadas que forem mais bem sintatica e semanticamente definidas, ou seja, que o processo de identificação apresentar certa "trivialidade", tais como nomes de pessoas, lugares, organizações, datas, etc, serão referenciadas apenas como *nomes próprios*.

Por sua vez, as entidades nomeadas que não pertencerem ao grupo de nomes próprios, e que não pertencerem a nenhuma classe específica de entidades, mas que seu significado for relevante ao ponto de poder ser ligada a uma entidade contida na base de entidades, serão classificadas como *conceitos concretos*.

O framework UnBWikilinks [1], apresentado no capítulo 4, explora de maneira específica o problema de ligação de entidades envolvendo conceitos concretos, e a definição de conceito concreto adotada nesse trabalho deriva diretamente da definição utilizada pelo framework.

3.7 Resumo do Capítulo

Neste capítulo foi apresentado o problema da ligação de entidades, que trata da identificação de referências em textos diversos a entidades como pessoas, lugares e organizações. Trata-se de um problema relevante da área de Processamento de Linguagem Natural (PLN) que tem despertado o interesse de diversos pesquisadores.

O capítulo apresentou, também, o framework comum para a realização de LE, consistindo de três passos principais: detecção das menções (ou referências às entidades); classificação e seleção das entidades; e desambiguação/otimização dos resultados. Essas

etapas também são executadas neste trabalho, e os detalhes de cada fase são apresentados nos próximos capítulos.

A seção 3.3 apresentou as principais bases de entidades abordadas pelos estudos mais recentes. As bases de entidades são componentes fundamentais do processo de realização de LE, e o sucesso de uma ferramenta está diretamente relacionado à qualidade da base. Neste trabalho foi adotada uma base de entidades formadas por páginas da Wikipédia.

Capítulo 4

Estado da Arte

Neste capítulo são apresentados os principais estudos na área de ligação de entidades. A seção 4.1 introduz o framework básico para a ligação de entidades, utilizado na comparação de algumas ferramentas. A seção 4.2 apresenta uma abordagem baseada em Máquinas de Vetores de Suporte (MVS). Já a seção 4.3 apresenta uma abordagem que emprega Modelos de Espaço Vetorial (MEV). As seções 4.4, 4.5, 4.6, 4.7, 4.8 e 4.9 apresentam, respectivamente as ferramentas Aida, Wikify!, DBpedia Spotlight, TagMe, Wikipedia Miner e LINDEN. A seção 4.10 apresenta em detalhes o estudo que deu origem ao UnBWikilinks e que serviu de inspiração para o presente trabalho. Por fim, a seção 4.11 traz um resumo do conteúdo apresentado no capítulo.

4.1 Um Framework para Ligação de Entidades

A comparação de diferentes abordagens pode ser realizada por meio de um framework para a realização da ligação de entidades (LE). A principal tarefa de um sistema de LE é vincular uma determinada menção com sua respectiva entidade em uma base de entidades. Essa tarefa pode ser dividida em três componentes principais [15]:

- *Extração* - é a detecção e a preparação das menções existentes em um dado texto. A maioria das bases de dados dos sistemas de LE fornecem strings para as menções por meio de consultas. Algum tratamento adicional nas menções pode ser desejável pois outras informações obtidas no texto podem ser úteis para desambiguação. A fase de extração pode incluir, também, tokenização, detecção de limites dos termos, e correferência em documentos. Correferência, em particular, é importante, pois pode ser usada para encontrar termos de pesquisas mais específicos (por exemplo, ABC - Australian Broadcasting Corporation).

- *Busca* - é o processo de geração de um conjunto de entidades candidatas, da base de entidades, para cada menção do texto. Títulos e outras características de páginas da Wikipédia, por exemplo, podem ser aproveitadas nesta fase para capturar sinônimos.
- *Desambiguação* - é o processo por meio do qual a melhor entidade é selecionada, da lista de entidades candidatas, para cada menção. Esta etapa pode ser entendida como um problema de classificação.

Nas seções a seguir serão apresentadas informações sobre as principais pesquisas na área, levando sempre em consideração o framework ora citado.

4.2 Utilizando MVS

Bunescu e Pasca (2006) foram os primeiros a explorar a tarefa de LE utilizando Máquinas de Vetores de Suporte (MVS) para a classificação em tarefas de desambiguação. Os componentes foram empregados da seguinte forma [3]:

- Extrator - foram utilizados dados derivados da Wikipédia como avaliação cujo objetivo é retornar o alvo correto para um determinado link, ou seja, reintroduzir links alvos em páginas da Wikipédia para determinadas âncoras do texto. Não foram executados pré-processamentos adicionais.
- Busca - o componente de pesquisa utiliza correspondência exata com os títulos dos artigos, páginas de redirecionamento e páginas de desambiguação. Os artigos candidatos correspondentes são retornados.
- Desambiguação - o componente de desambiguação adotado implementa MVS como modelo de classificação, tendo sido empregada a ferramenta SVM$^{\text{light}}$ [18]. Duas características são usadas. A primeira característica é a similaridade entre o contexto da consulta e as páginas candidatas, calculada através da função cosseno a seguir; a segunda característica é a criação de uma 2-tupla para cada combinação de categorias de candidatos - categorias da Wikipédia que são usadas para agrupar objetos semelhantes - e palavras de contexto:

$$\cos(q,d) = \frac{\|T_q \cap T_d\|}{\max\|T_q \cap T_d\|} \times \sum_{t \in T_d} \sqrt{tf(t,d)} \times (1 + \log \frac{\|D\|}{df(t)}) \times \frac{1}{\sqrt{\|T_d\|}} \qquad (4.1)$$

Na função acima, q é o texto do contexto da consulta, d é documento, T_i é o conjunto de termos em i, M é o conjunto de documentos que correspondem à consulta q, *tf(t,d)* é a frequência do termo t no documento d, D é o conjunto completo de documentos e *df(t)*

é a contagem de documentos em D que possuem o termo t. A ferramenta desenvolvida foca na identificação e LE de pessoas/personalidades.

4.3 Uma abordagem MEV

Cucerzan (2007) descreve uma abordagem LE que se concentra em desambiguação no nível de documento. Ele também apresenta um módulo de pré-processamento que identifica cadeias de entidades correferentes a fim de utilizar sequências de nomes mais específicos para a consulta. No entanto, o efeito da manipulação da correferência na busca e na desambiguação não é explorado [8].

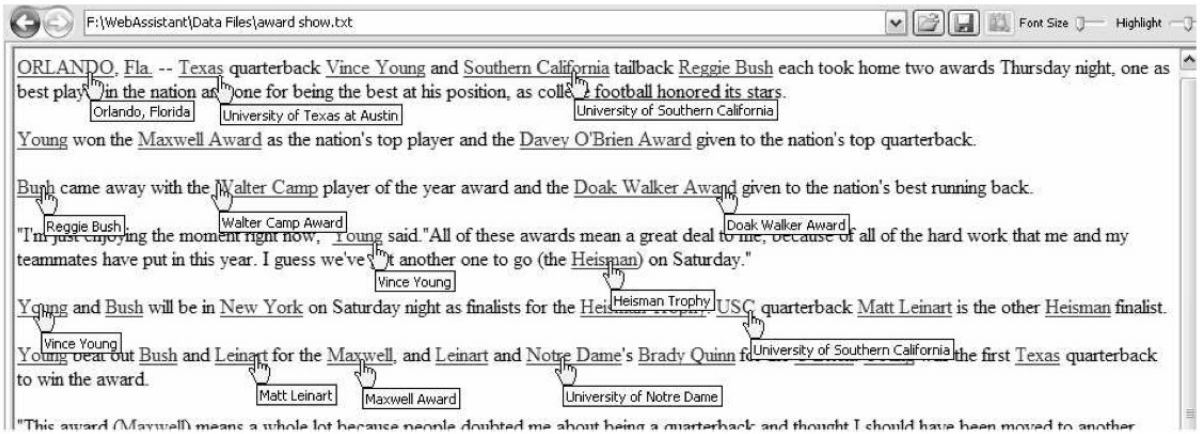

Figura 4.1: Exemplo de resultado do processamento proposto por Cucerzan. Fonte original em inglês: Cucerzan [8].

- Extrator - o objetivo é ligar todas as entidade identificadas em notícias com suas correspondentes página da Wikipédia. Portanto, é necessário dividir o texto em frases para, em seguida, realizar a ligação das entidades. Foi utilizado um Recuperador de Entidades Nomeadas (NER) híbrido baseado em regras de capitalização, web e estatísticas dos dados compartilhados do CoNLL-03 [50].

- Busca - para geração das entidades candidatas, as menções canônicas são primeiramente normalizadas para se adequarem às convenções da Wikipédia. Elas são pesquisadas por meio de correspondência exata com os títulos dos artigos, páginas de redirecionamento e páginas de desambiguação.

- Desambiguação - as consultas contendo as menções são desambiguadas em relação aos vetores-documento derivados de todas as entidades. São construídos vetores representando os documentos e todas as entidades candidatas, cada candidata de cada

entidade canônica. Um vetor com variáveis indicadoras dos candidatos é criado para cada candidato, baseado na presença de categorias e contextos contidos no artigo. Os contextos são âncoras extraídas do primeiro parágrafo, ou aqueles que conectam o artigo a um outro e vice-versa. O vetor documento estendido é preenchido para representar a união das variáveis indicadoras de todos os vetores de entidades. Os valores das categorias são a quantidade de vetores-entidade que contém a categoria e o contexto correspondente ao contexto do documento. Cada lista de candidatas para cada menção é classificada separadamente com relação ao vetor no nível do documento. Mais precisamente, as candidatas são classificadas conforme o score obtido do produto escalar do vetor candidato e do vetor documento estendido, com penalidade para evitar dupla contagem. Um exemplo do resultado do processamento é exibido na figura 4.1.

4.4 AIDA: Desambiguação precisa

Apresentado em 2011, o framework AIDA caracteriza-se como uma ferramenta *online* para realizar identificação e desambiguação de entidades [56]. Dado um texto escrito em linguagem natural, a ferramenta mapeia menções relacionadas a nomes ambíguos para entidades, tais como pessoas e lugares, registradas em bases de conhecimento como DBpedia, Freebase ou YAGO [49].

A entrada para AIDA é um texto arbitrário, opcionalmente em HTML ou XML, com menções de entidades nomeadas (pessoas, bandas de música, canções, universidades, etc.). As menções são detectadas automaticamente usando o *Stanford NER Tagger*. Para o mapeamento coletivo é utilizada uma abordagem baseada em grafo. O grafo é construído com menções e as suas entidades candidatas como nós do grafo. Existem dois tipos de relacionamentos:

- *menção-entidade*: entre menções e suas entidades candidatas com pesos que capturam a similaridade entre o contexto de uma menção e de uma entidade candidata; e
- *entidade-entidade*: entre diferentes entidades com pesos que capturam a semelhança semântica entre duas entidades.

O objetivo é reduzir o grafo a um sub-grafo denso, onde cada nó-menção está ligado a um e apenas um nó-entidade candidata, que fornece a saída do mapeamento. A densidade refere-se ao peso total das extremidades do sub-grafo, ou, alternativamente, ao grau mínimo ponderado no sub-grafo. Uma vez que o grafo é construído, um algoritmo guloso é utilizado para calcular o sub-grafo. Em cada iteração são realizadas duas etapas:

- identificar o nó-entidade que tem o menor grau ponderado; e
- remover esse nó e suas bordas incidentes do grafo, a menos que seja a última entidade candidata restante para uma das menções.

A figura 4.2 ilustra o grafo menção-entidade para um texto de entrada com as menções destacadas (à esquerda) e entidades candidatas (centro) conforme o banco de dados (à direita). A espessura das arestas entre as entidades retrata os diferentes pesos das arestas.

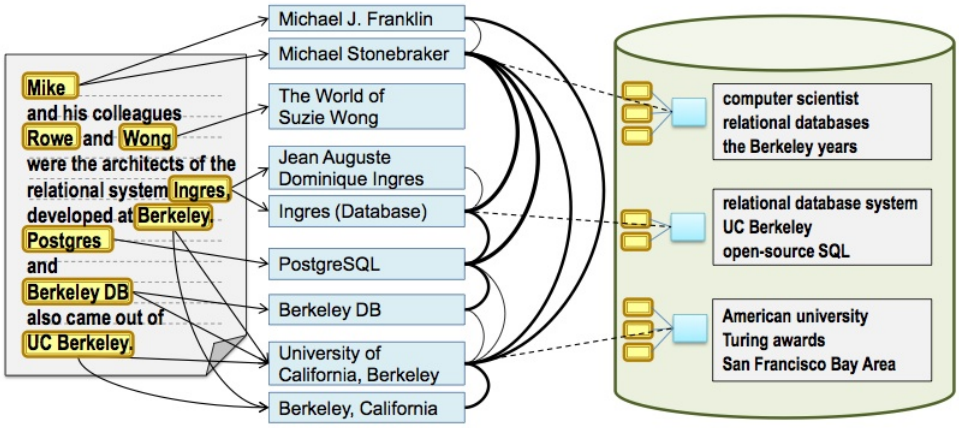

Figura 4.2: Exemplo do grafo menção-entidade. Fonte original em inglês: Yosef *et al.* [56].

A semelhança entre uma menção e uma entidade candidata é calculada como uma combinação linear de dois ingredientes. O primeiro é o destaque de uma entidade, por exemplo, o príncipe "Harry" da Inglaterra *versus* "Harry Kelly", um jogador de basquete americano menos conhecido. Esse ingrediente funciona como uma probabilidade prévia para cada mapeamento potencial. Essa prévia é calculada com base nas estatísticas sobre âncoras *href* e seus alvos de links na Wikipédia. O segundo ingrediente para os pesos das arestas é baseado na sobreposição entre o contexto de uma menção e o contexto de uma entidade candidata. Para uma menção é considerado o texto de entrada completo como seu contexto. Para as entidades são consideradas palavras-chave das entidades, pré-computadas a partir dos artigos da Wikipédia que as entidades da base YAGO armazena. Tais palavras-chave são todas as frases em âncoras de links, incluindo nomes de categorias, títulos de citações e referências externas no artigo da entidade.

Em entidades com texto longo pode ocorrer queda na precisão devido à pequena quantidade de correspondências de palavras. Para evitar essa queda na precisão é considerado o tamanho da janela que comporta todas as palavras-chave que aparecem no texto de entrada. Além disso, palavras-chave podem ser penalizadas em função da distância em relação à menção em questão. Uma vez que são realizadas comparações parciais, palavras

diferentes em uma frase têm diferentes graus de importância. Esse aspecto é tratado com a coleta de estatísticas de grandes bases (por exemplo, Wikipédia) sobre a frequência de co-ocorrência de uma palavra de uma entidade de interesse.

As entidades candidatas são obtidas a partir da base de conhecimento YAGO [49] e o foco é a identificação do que chamamos nesse trabalho de nomes próprios, ou seja, sentenças que indicam pessoas, lugares, personalidades, empresas, etc., e que são mais bem sintaticamente definidas.

4.5 Wikify!: Um sistema para identificação de entidades nomeadas

O Wikify! introduz o uso da Wikipédia como um recurso para a extração automática de palavras-chave e desambiguação de palavras. Especificamente, para um dado documento de entrada, a ferramenta tem a capacidade de identificar os conceitos importantes de um texto (extração de palavras-chave) e, em seguida, vincular esses conceitos com as páginas da Wikipédia correspondentes (desambiguação) [33].

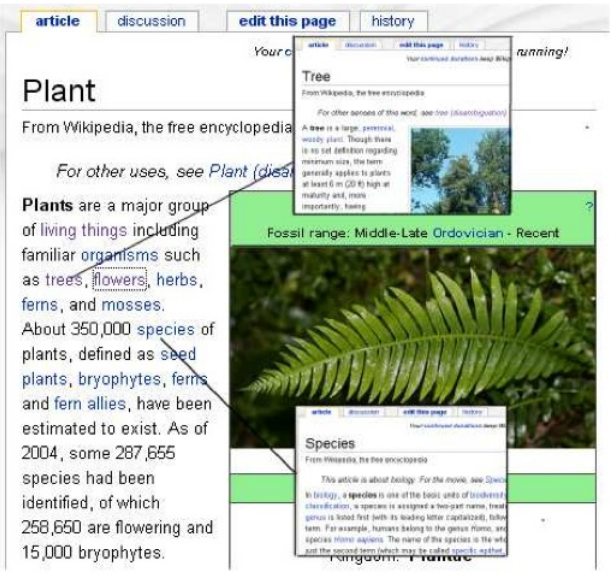

Figura 4.3: Exemplo de página da Wikipédia com os links relacionados aos artigos. Fonte original em inglês: Mihalcea e Csomai [33].

Dado um documento de texto, *wikificação* é definida como a tarefa de extrair automaticamente as palavras e frases mais importantes no documento identificando, para cada palavra-chave, um artigo apropriado da Wikipédia. Esta é a tarefa típica realizada pelos usuários da Wikipédia ao contribuir com artigos para o repositório da enciclopédia.

O objetivo é adicionar links para os conceitos mais importantes em um documento, permitindo que os leitores sigam diretamente a linha de raciocínio para outros artigos. Em geral, as ligações representam "grandes conexões com o tema de outro artigo que auxiliará os leitores a entenderem o artigo atual mais plenamente". Um exemplo dessa ligação é exibido na figura 4.3 que apresenta um texto sobre "plantas" e links para dois conceitos importantes destacados no texto ("árvores" e "espécies") que podem ampliar a compreensão do leitor sobre o assunto abordado.

A primeira tarefa consiste em identificar palavras e frases que são consideradas importantes para o documento. Geralmente incluem termos técnicos, novas terminologias, bem como outros conceitos intimamente relacionados com o conteúdo do artigo - em geral, todas as palavras e frases que irão acrescentar conteúdo à experiência do leitor.

A segunda tarefa consiste em encontrar o artigo Wikipédia correto que deve ser ligado à cada palavra-chave selecionada. Aqui, o problema da desambiguação é tratado já que uma frase pode ser geralmente ligada a mais de uma página da Wikipédia, e a interpretação correta da frase (e, correspondentemente, o link correto) depende do contexto em que ela ocorre.

A tarefa de desambiguação é realizada utilizando a *Normalized Google similarity distance*, uma medida de similaridade semântica aplicada para dimensionar o relacionamento entre páginas da Wikipédia. O framework foi desenvolvido para tratar textos em inglês de tamanhos variados, e o software encontra-se disponível para download [7].

4.6 DBpedia Spotlight: Ontologia DBpedia

Em 2011 foi apresentado o DBpedia Spotlight [32] um framework para anotação de textos com entidades da DBpedia que possibilita aos usuários configurar as anotações para suas necessidades específicas através da ontologia DBpedia. O maior desafio está justamente na desambiguação, e a estratégia adotada para superar essa dificuldade é a utilização da DBpedia: uma base de conhecimento formada a partir da Wikipédia, estruturada em uma ontologia com classes tais como pessoas, organizações, atletas, empresários, etc.

Uma das inovações do DBpedia Spotlight é justamente ampliar as classes de menções identificadas uma vez que o framework se propõem anotar qualquer uma das 272 classes disponíveis na ontologia, enquanto as demais ferramentas estão restritas às classes de palavras mais triviais. Na figura 4.4 é exibida a interface da ferramenta com alguns exemplos das classes de entidades que podem ser anotadas pelo framework.

A base de conhecimento é formada por links, redirecionamentos e páginas de desambiguação contidas na Wikipédia, compilada a partir do pré-processamento dos artigos da

enciclopédia que mapeou os links de maneira que cada âncora representa uma forma de escrita da menção e o link propriamente dito representa uma fonte.

Um outro diferencial do framework é justamente o uso de representação baseada em Modelos de Espaço Vetorial, onde a medida de similaridade para classificação dos vetores é o cosseno. O cálculo leva em consideração a relevância de uma palavra para uma fonte (frequência de termos ou TF) e a habilidade de identificar uma dada forma de escrita para uma palavra entre diversas candidatas (frequência inversa ou ICF).

Figura 4.4: DBpedia Spotlight: interface web. Fonte original em inglês: Mendes *et al.* [32].

A avaliação da ferramenta foi realizada comparando os resultados obtidos no tratamento de textos extraídos da própria Wikipédia e artigos do jornal *New York Times* em relação aos demais serviços disponíveis publicamente de anotação de entidades, tendo apresentado os melhores resultados nas comparações efetuadas.

4.7 TagMe: Anotação de pequenos fragmentos de textos

O sistema TagMe [10] tem como objetivo identificar em textos passados como entrada os correspondentes links para páginas da Wikipédia. Por focar em pequenos fragmentos de texto, o estudo envolvendo a implementação do TagMe teve que lidar com a dificuldade de anotar textos com pouca informação assessória, tais como tweets, manchetes, etc., dificultando sobremaneira a tarefa de desambiguação.

TagMe usa âncoras da Wikipédia e as páginas relacionadas com elas como os seus possíveis sentidos no processo de desambiguação, resolvendo a ambiguidade e a polissemia,

utilizando potencialmente os mapeamentos de âncora-página disponíveis, encontrando a concordância entre eles por meio de novas funções de pontuação que são rapidamente computadas, eficazes na anotação produzida.

De maneira mais detalhada, na abordagem proposta cada menção é associada com um conjunto de entidades candidatas, e a desambiguação é realizada levando-se em consideração a estrutura de grafo da Wikipédia, de acordo com a métrica de parentesco [34] que considera a quantidade de links comuns entre duas entidades. Além disso, um conjunto de heurísticas é eventualmente adotado para selecionar a melhor anotação para cada menção.

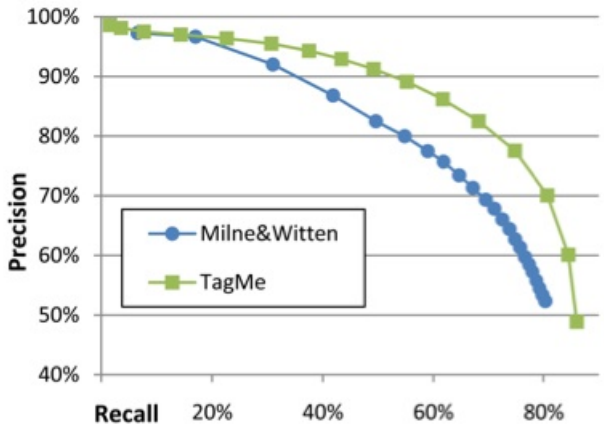

Figura 4.5: Comparação do TagMe com o estado da arte. Fonte original em inglês: Ferragina e Scaiella [10].

Experimentos preliminares mostram que TagMe supera os sistemas contemporâneos quando eles são adaptados para trabalhar com textos curtos, conforme gráfico exibido na figura 4.5. Adicionalmente, TagMe obtém resultado competitivo com textos longos também. O framework TagMe possui API disponível publicamente de forma online.

4.8 Wikipedia Miner: *Wikification*

O trabalho relacionado ao framework Wikipedia Miner pode ser dividido em duas tarefas principais. A primeira delas concentra-se na desambiguação, onde foi adotada uma abordagem de aprendizado de máquina que utiliza os links encontrados dentro dos artigos da Wikipédia para o treinamento. Para cada link, um humano, manualmente, e provavelmente com algum esforço, selecionou o destino correto para representar o sentido pretendido dos links. Existem milhões de exemplos manualmente definidos para serem usados no aprendizado [34].

A abordagem tem como princípio equilibrar a frequência de um sentido com a sua ligação ao contexto do texto. A frequência de um sentido é definida pelo número de vezes

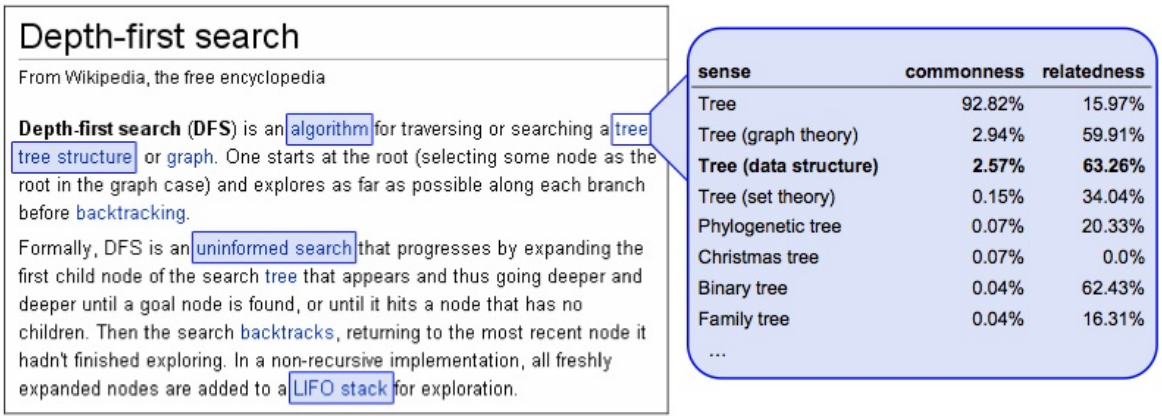

Figura 4.6: Desambiguação da palavra *tree* usando links não ambíguos como contexto. Fonte original em inglês: Milne e Witten [34].

em que foi utilizado como um destino na Wikipédia. Nem sempre a mais alta frequência é a melhor decisão, conforme figura 4.6. O algoritmo proposto identifica esses casos, comparando cada sentido possível com o seu contexto envolvente.

A escolha acima leva em consideração uma medida sugerida pela abordagem, conhecida como *relatedness*, que representa a média ponderada da correlação de um sentido com cada contexto de cada artigo, e que pode ser obtida por meio da seguinte equação:

$$relatedness(a,b) = \frac{\log(max(\|A,B\|)) - \log(\|A \cap B\|)}{\log(\|W\|) - \log(min(\|A,B\|))} \tag{4.2}$$

A segunda tarefa consiste na detecção dos links. Para realizar esta tarefa é sugerido um classificador que considera três características diferentes para obter anotações válidas e descartar anotações irrelevantes: a probabilidade prévia de uma menção se referir a uma entida de específica; o relacionamento com o contexto em que a entidade é extraída; e a qualidade do contexto que leva em consideração o número de termos envolvidos, a extensão com que eles se relacionam e a frequência com que eles são utilizados na Wikipédia.

Projetado para lidar com documentos de qualquer tamanho em inglês, o framework possui API disponível de forma online.

4.9 LINDEN: Novas classes de entidades

O framework LINDEN explora de maneira objetiva a tarefa de ligação de entidades [46], investigando a necessidade de se construir bases de conhecimentos cada vez mais abrangentes e atualizadas (entidades, respectivas classes semânticas e relacionamentos mútuos). O objetivo é selecionar a entidade mais adequada para cada menção extraída do texto, ao invés de listar um conjunto de entidades forçando o usuário a tentar aprimorar

sua pesquisa, e retornar *vazio* quando nenhuma entidade candidata for suficientemente significativa.

Na abordagem proposta pelo LINDEN, que utiliza tanto a Wikipédia como a base YAGO como fontes de conhecimento, para cada entidade nomeada um conjunto de entidades candidatas é selecionado com base em quatro características da Wikipédia: páginas, redirecionamentos, páginas de desambiguação e hiperlinks. Essa representação é exemplificada na figura 4.7.

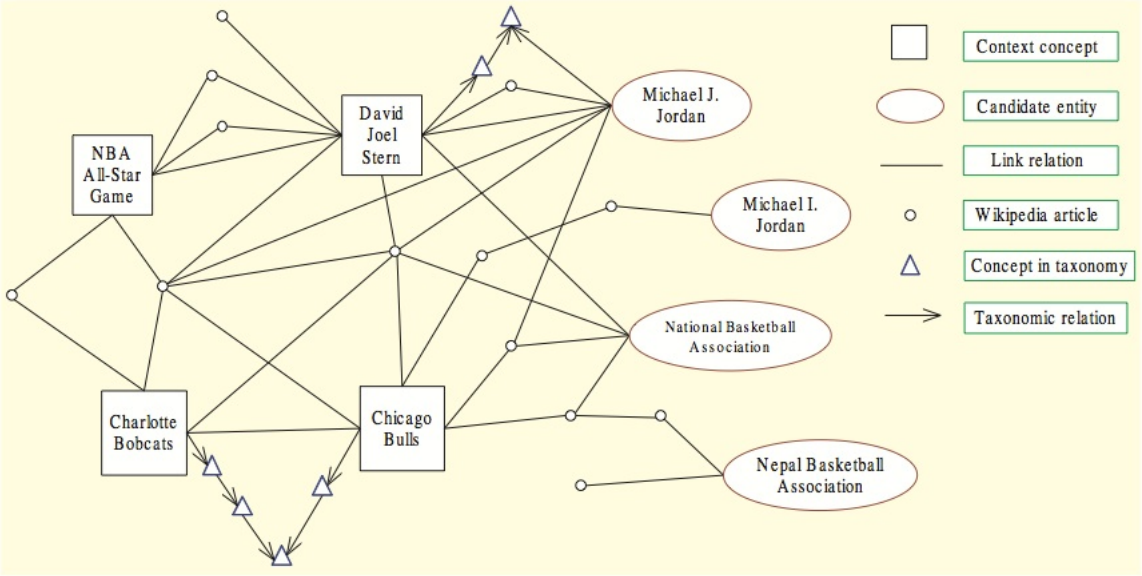

Figura 4.7: LINDEN: exemplo de construção de rede semântica. Fonte original em inglês: Shen *et al.* [46].

Em seguida, ocorre o processo de desambiguação para selecionar a entidade que melhor se adeque ao contexto. Nesse processo, cada entidade recebe uma nota baseada nos parâmetros probabilidade do link, associatividade semântica, similaridade semântica e coerência global. No cálculo desses parâmetros, a ferramenta utiliza o Wikipedia Miner para a identificação de conceitos, a *Wikipedia Link-Based Measure* para calcular a associatividade semântica de cada entidade candidata, a base de dados YAGO para identificar as variadas classes de entidades existentes e, por fim, a coerência global que relaciona a associatividade semântica de uma entidade com as demais.

A eliminação das entidades não representativas é determinada quando a nota calculada não atinge uma nota de corte mínima. Nessas situações, a ferramenta retorna *vazio* evitando que uma entidade candidata seja atribuída indevidamente a uma entidade nomeada extraída do texto.

4.10 UnBWikilinks: Ligando conceitos concretos

O framework UnBWikilinks, desenvolvido na Universidade de Brasília como solução para o desafio apresentado na edição de 2013 do *Web Information System Engineering* (Wise 2013), tem como principal objetivo realizar a ligação de entidades envolvendo conceitos concretos. A organização do Wise 2013 exemplificou conceito concreto como sendo classes de palavras que não apresentam conceitos comuns, triviais, tais como "hotel", "viagem", "livro", etc., mas sim entidades que trazem consigo uma maior carga de significado, tais como "agências de viagens online", "caracteres chineses", etc.

4.10.1 Análise dos Dados

Para a solução do desafio os organizadores do Wise 2013 forneceram uma base de Wikilinks, que possui mais de 40 milhões de menções referentes a aproximadamente 3 milhões de entidades. Trata-se de uma base contendo hiperlinks extraídos da web para páginas da Wikipédia em Inglês. Na base fornecida foram identificados 2,8 milhões de entidades e aproximadamente 5,8 milhões de menções.

Após a tokenização do arquivo contendo os wikilinks, foram encontrados aproximadamente 8,2 milhões de tokens (unidades de texto, tais como palavras e quantidades), com distribuição irregular entre as entidades. A média de tokens por entidade é igual a 2; o máximo número de tokens em uma mesma entidade é igual a 72; e o número total de tokens nas menções é de aproximadamente 21,7 milhões (mínimo de 3 e máximo de 588 tokens por entidade).

4.10.2 Descrição Formal dos Dados

A formalização adotada seguiu as abordagens tradicionais para o problema de "desambiguação com a Wikipédia" (D2W). Uma ilustração dos termos segue na figura 4.8, exemplificando o processo de análise:

A figura 4.8 apresenta um documento de texto d contendo um conjunto de entidades nomeadas (EN) $N = n_1, ..., n_N$. O termo entidade nomeada foi utilizado para indicar a ocorrência de um nome próprio ou de um conceito concreto dentro de um texto simples.

O objetivo é produzir um mapeamento a partir do conjunto de entidades nomeadas para o conjunto de entidades da Wikipédia (URLs) $W = \{e_1, ..., e_{|W|}\}$ extraídas de um arquivo f contendo uma lista de entidades e sua menções, onde cada linha representa uma URL da Wikipédia. É possível que uma EN corresponda a uma entidade que não esteja listada em f, portanto, uma entidade nula é adicionada ao conjunto W. Cada entidade da Wikipédia tem um conjunto de menções $M = \{m_1, ..., m_n\}$ associado no arquivo f.

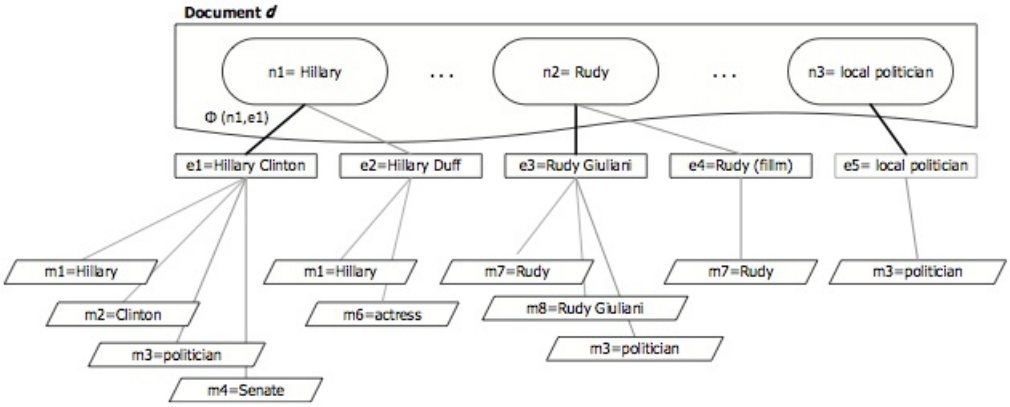

Figura 4.8: Representação do processo de análise. Fonte original em inglês: Abreu *et al.* [1].

O problema pode ser expresso como sendo o de encontrar uma correspondência um-para-muitos em um grafo bipartido, com EN formando uma partição e as entidades da Wikipédia outra partição (figura acima). A correspondência de saída é denotada como uma *N-tupla* $\gamma = (e_1, ..., e_N)$, onde e_i é a correspondência mais precisa, ou a desambiguação, para a EN n_i.

A correspondência é definida por alguns parâmetros. Supondo $\phi(n_i, e_j)$ como sendo uma função de pontuação que reflete a probabilidade de que a entidade $e_j \in W$ é a correlação mais precisa para $n_i \in N$. Pesquisas anteriores consideram que identificar todas as entidades nomeadas com base em uma abordagem local pode ser expressa como um problema de otimização, da seguinte forma:

$$\gamma_{local} = argmax_r \sum_{i=1}^{N} \phi(n_i, e_j) \tag{4.3}$$

Abordagens locais definem uma função ϕ para estabelecer qual é a combinação mais precisa, atribuindo pontuações mais altas para entidades com conteúdo similar ao do documento de entrada. Abordagens globais funcionam de uma maneira mais complexa, combinando todo o conjunto de EN simultaneamente, com o objetivo de melhorar a coerência entre as entidades ligadas.

Mesmo com uma abordagem mais simples, como correspondência local, os resultados obtidos podem ser competitivos com as abordagens globais tradicionais, ou seja, é possível chegar a resultados muito interessantes com recursos limitados.

4.10.3 Proposta de Solução

A abordagem apresentada pelo UnBWikilinks é baseada em cinco etapas:

1. pré-processamento do arquivo contendo as entidades, resolvendo questões relacionadas à formatação;
2. tokenização e análise POS dos textos passados como entrada;
3. processamento dos textos com o objetivo de detectar nomes próprios e conceitos concretos segundo uma série de regras;
4. identificação da respectiva entidade para cada nome próprio/conceito concreto; e
5. compilação do resultado na forma de arquivo, listando todas as menções e as respectivas entidades [1].

No primeiro passo, o arquivo de entidades (wikilinks) é processado e, como resultado, as tabelas da base de dados são criadas, conforme ilustrado na figura 4.9. Todas as entidades e suas respectivas menções são organizadas em duas tabelas separadas, contendo referências para o arquivo original. Cada entidade é processada para criar uma tabela auxiliar que contém os tokens dos núcleos das palavras da entidade, que serão utilizados como parâmetro de combinação.

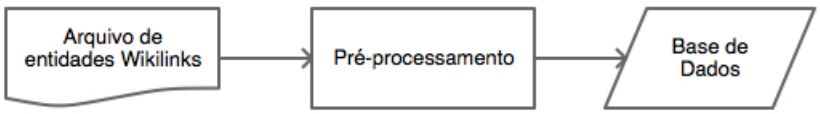

Figura 4.9: Passos executados no processamento do arquivo de entidades. Fonte: adaptado de Abreu *et al.* [1].

Uma ferramenta de taggeamento automático anota cada menção, e para cada token é atribuída uma classificação *part-of-speech* (POS) de acordo com o conjunto de tags do Penn Treebank. O resultado do POS é mantido no banco de dados e será utilizado na análise e na extração das sentenças.

Algumas estatísticas são calculadas na etapa de pré-processamento: o número de menções para cada entidade e o número de vezes que um token núcleo de uma palavra está listado nas menções das entidades. Essas estatísticas também serão usadas como entrada para uma função de pontuação na etapa correspondente.

Para cada texto do conjunto de testes, algumas etapas de pré-processamento também são necessárias para evitar problemas de formatação de baixo nível. Como é impossível prever a qualidade da fonte do conjunto de testes, podem haver vários padrões de formatação não triviais para a solução e, consequentemente, precisam ser filtrados. Em

particular, a solução converte todas as primeiras palavras de uma frase para iniciais minúsculas. Iniciais maiúsculas encontradas no meio de uma frase são mantidas como estão.

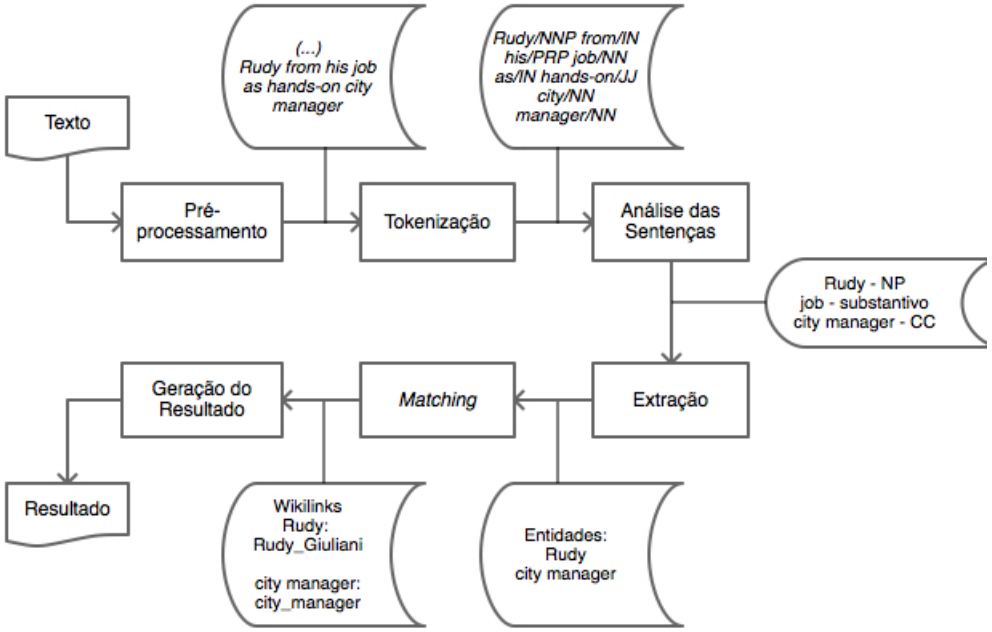

Figura 4.10: *Overview* da arquitetura da ferramenta UnBWikilinks. Fonte: adaptado de Abreu *et al.* [1].

Os textos pré-processados são divididos em tokens e passam por uma análise POS para receber a tag apropriada. O processo de geração de tokens é um primeiro passo comum em extração de informação (EI). O conjunto de dados Penn Treebank foi adotado por suas especificidades, por exemplo, por apresentar as diferenças entre substantivos comuns e próprios.

O uso de tokens é feito para o texto como um todo. Trabalhos anteriores comprovaram que é possível evitar o custo de analisar os documentos completos. No entanto, para uma melhor precisão todos os textos foram marcados por inteiro.

A etapa de análise das sentenças visa identificar grupos de nomes próprios e conceitos concretos. O sistema só realiza uma análise parcial: construir a estrutura que o Wise 2013 requereu. Ao contrário das ferramentas tradicionais, um analisador parcial procura por fragmentos de texto que podem ser reconhecidos de forma confiável.

A estrutura gramatical de nomes próprios e conceitos concretos foi analisada, e desse trabalho foi proposto um par de expressões regulares com o objetivo de orientar a solução a encontrar tais termos dentro do texto simples. Esta técnica identifica tais fragmentos de

maneira determinística com base em dicas sintáticas puramente locais. Por este motivo, a sua cobertura é limitada.

Um grupo de expressões regulares foi definido com base nos níveis do conjunto de tags Penn Treebank POS em Inglês. A maioria das ocorrências de nomes próprios e conceitos concretos são restritas a 5 tags: / JJ adjetivo, / NN substantivo no singular, / NNS substantivo no plural, / NP nome próprio no singular e / NPS nome próprio no plural. A descrição das ocorrências é mostrada abaixo:

- Nomes Próprios: $p + n$?
- Conceitos Concretos: a ? $n +$

Nas regras acima, p indica substantivos próprios, n indica substantivos comuns e a indica adjetivos. Os qualificadores (?) e (+) indicam zero ou um e um ou mais elementos, respectivamente. A figura 4.11 apresenta um exemplo da tarefa exigida no Wise 2013:

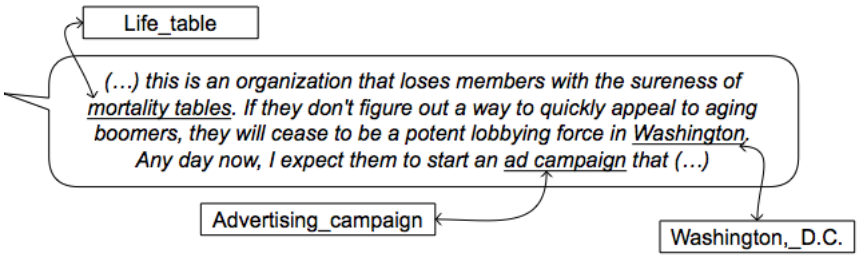

Figura 4.11: Exemplo de LE no Wise 2013.

O resultado da análise de cada sentença é um conjunto de EN extraídas do texto simples que correspondam a um nome próprio ou a um conceito concreto. A EN, tal como é definida pelas expressões regulares, pode ser composta por uma ou mais palavras. A extração da entidade é a ação de identificar o conjunto de termos da sentença que pode ser de potencial relevância. A arquitetura da metodologia UnBWikilinks é exibida na figura 4.10.

O objetivo de buscar a entidade correspondente é determinar quais EN referem-se à entidades wikilinks. Para atingir esse objetivo, é definida uma função de pontuação $\phi(n_i, e_j)$ para refletir a probabilidade de que a entidade $e_j \in W$ é a mais adequada para $n_i \in N$. Boa parte dos sistemas usam heurísticas geradas manualmente para determinar quando duas frases descrevem a mesma entidade, mas gerar boas heurística que cubram todos os tipos de referências ainda é um desafio.

Após a extração da EN, o algoritmo pesquisa no banco de dados por todas as entidades que são semelhantes à EN. O próximo passo é calcular a função de pontuação $\phi(n_i, e_j)$.

São calculados 4 parâmetros, variando de 0 a 1, sendo 1 a combinação perfeita para os parâmetros. Os quatro parâmetros da solução são:

A : Correspondência com as palavras dos títulos dos artigos.

B : Correspondência com o contexto da frase.

C : Número de menções na entidade.

D : Correspondência com o contexto do texto.

A função score é definida conforme segue:

$$\phi(n_i, e_j) = \alpha A + \beta B + \gamma C + \theta D \tag{4.4}$$

O parâmetro C independe da palavra recebida e refere-se ao tamanho da entidade. Todas as constantes $\alpha, \beta, \gamma, \theta$ foram calibradas manualmente [1].

Os participantes do Wise 2013 foram convidados a processar um conjunto contendo aproximadamente 8.800 textos e extrair nomes próprios e conceitos concretos com as respectivas entidades da base Wikilinks. Os resultados enviados pelos participantes foram comparados com um gabarito oficial, e foram selecionados três trabalhos que obtiveram as melhores avaliações, incluindo a metodologia UnBWikilinks que obteve a maior *precisão* na análise combinada de nomes próprios e conceitos concretos, como pode ser verificado na tabela 4.1. Os resultados dos dois outros trabalhos selecionados, identificados como *Equipe 299* e *Equipe 306*, também estão exibidos na tabela 4.1.

Tabela 4.1: Resultados obtidos pela ferramenta UnBWikilinks no Wise 2013. Fonte: adaptado de Chen *et al.* [5].

Trabalho Avaliado	***Recall* NP + CC**	***Recall* NP**	***Precision* NP + CC**
UnBWikilinks	40,1%	38,7%	42,5%
Equipe 299	47,5%	44,9%	14,0%
Equipe 306	44,1%	45,6%	27,8%

O presente trabalho pretende aprofundar no estudo da ligação de entidades envolvendo conceitos concretos, extendendo a pesquisa que deu origem ao UnBWikilinks. Na nova abordagem, a representação baseada em Modelo de Espaço Vetorial foi empregada com o objetivo de explorar, de maneira mais satisfatória, a análise do contexto em que as menções candidatas a conceitos concretos estão inseridas.

4.11 Resumo do Capítulo

Neste capítulo foram apresentados os mais recentes estudos na área de ligação de entidades. Inicialmente foram apresentados os três componentes principais do processo de LE: extração, busca e desambiguação. Tais componentes estão presentes em grande

parte dos trabalhos apresentados neste capítulo, sendo justamente a forma de aplicação de cada componente o principal diferencial entre as abordagens investigadas.

As pesquisas apresentadas neste capítulo compartilham alguns desses componentes. Os extratores são comumente construídos com base em dados da Wikipédia, estatísticas de diferentes bases, heurísticas, análise gramatical, dentre outras. Já a busca é baseada, em sua grande maioria, na correspondência das menções com as palavras contidas nas entidades. Por fim, a desambiguação, componente que apresenta maior particularidade, é realizada com o emprego de diferentes técnicas, tais como Máquinas de Vetores de Suporte, Modelos de Espaço Vetorial, grafos, métricas próprias (*Normalized Google similarity distance* e *relatedness*, por exemplo), dentre outras.

Por fim, o capítulo apresentou o estudo dedicado à ferramenta UnBWikilinks, framework pioneiro na ligação de conceitos concretos com entidades da Wikipédia. Alguns dos assuntos abordados nesse estudo são retomados no presente trabalho.

Capítulo 5

Abordagem MEV para Ligação de Conceitos Concretos

Neste capítulo a metodologia proposta é apresentada em detalhes. O framework com suas respectivas etapas é descrito na seção 5.1 e a arquitetura é descrita na seção 5.2. Por fim, a seção 5.3 traz um resumo do conteúdo apresentado no capítulo.

5.1 Descrição da Metodologia

A abordagem ora sugerida para a ligação de entidades (LE) envolvendo conceitos concretos (CC) com entidades Wiki segue um framework comum também utilizado, mesmo que de maneira implícita, pelos demais pesquisadores na resolução de problemas de LE e desafios afins, conforme descrito no capítulo 3.

Esse framework consiste, basicamente, das seguintes etapas:

1. Detectar as menções (sentenças "ligáveis");
2. Classificar e selecionar as respectivas entidades; e
3. Desambiguar/melhorar os resultados com base no contexto (semântica).

No framework acima descrito o primeiro passo consiste em analisar textos passados como entrada com o objetivo de identificar sentenças que possam representar conceitos importantes, e que serão objeto das etapas seguintes. Trata-se de um passo de fundamental importância visto que o resultado da LE depende de um processo de identificação de sentenças relevantes bem executado. Por mais que o módulo de desambiguação seja eficiente, de nada adiantará se as sentenças candidatas para a LE corretas não forem devidamente identificadas.

O segundo passo tem como objetivo identificar, na base de entidades, as respectivas candidatas para cada sentença "ligável" selecionada no passo anterior. Este passo depende diretamente de como a base de entidades foi formada, ou seja, de quais características de cada entidade foram armazenadas no banco. Como resultado deste passo são geradas listas de entidades candidatas, uma para cada sentença, que serão objeto de desambiguação no passo seguinte.

Por fim, no terceiro passo ocorre a desambiguação das entidades. O objetivo aqui é analisar, com base em diferentes técnicas a depender da metodologia empregada, diversas características tanto da base de entidades como das sentenças ligáveis e dos textos de onde elas foram retiradas para decidir, entre duas ou mais entidades semelhantes, qual entidade é a mais representativa para cada sentença.

No capítulo 4 foi apresentado, também, um framework compilado por [15] semelhante ao roteiro comentado acima, porém em termos de três componentes principais: *extração*, que consiste na detecção e na preparação das menções existentes no texto analisado; *busca*, que consiste na geração do conjunto de entidades candidatas, obtidas a partir da base de entidades; e *desambiguação*, responsável pela seleção da melhor entidade, dentre as candidatas, para cada menção. Aqui o termo *menção* representa as *sentenças* do framework anterior.

É possível perceber várias semelhanças entre os dois processos e estabelecer, inclusive, um de-para direto: os passos 1, 2 e 3 referem-se aos componentes de *extração*, *busca* e *desambiguação* comentados no parágrafo anterior. Este trabalho também pode ser descrito nos mesmos termos. A metodologia ora proposta consiste, também, de três etapas principais:

1. Criação da base de entidades;
2. Extração de nomes próprios (NP) e conceitos concretos (CC); e
3. Ligação dos NP e CC com as entidades Wiki.

As etapas acima se relacionam com o framework básico da seguinte forma: a primeira etapa, de preparação da base de entidades, não é citada no framework talvez por ser intrínseca ao processo de LE, porém para fins de compreensão da metodologia será destacado, neste trabalho, como uma etapa particular e será detalhada nas subseções a seguir; a segunda etapa, como o nome sugere, se relaciona diretamente com o passo 1 do framework por tratar justamente da extração das sentenças/menções dos textos analisados; e a terceira etapa agrupa os passos 2 e 3 do framework básico, compreendendo as atividades de seleção, desambiguação e classificação das respectivas entidades.

A etapa de criação da base de entidades consiste no tratamento das bases da Wikipédia disponibilizadas de modo *offline*, para formar um banco de dados normalizado de

entidades e respectivas menções, entidades que serão ligadas posteriormente aos NP e CC extraídos dos textos.

Por sua vez, a etapa de extração de NP e de CC tem o objetivo de tratar o conjunto de textos passados como entrada com vistas a identificar NP e possíveis CC relevantes para a compreenção de cada texto. Apesar do foco ser a identificação de CC, a identificação de NP também foi mantida no escopo da metodologia.

Por fim, a etapa de ligação dos NP e CC com as entidades Wiki tem como objetivo identificar entidades candidatas da base de entidades e realizar a classificação/seleção. Tais entidades são representadas em um Modelo de Espaço Vetorial (MEV) para auxiliar na seleção da entidade com significado mais semelhante ao do NP e do CC tratado no momento.

As etapas acima podem ser expandidas para associação com o framework para qualquer sistema de extração de entidades sugerido por [4], que consiste dos cinco passos a seguir:

1. Os textos e a base de entidades são pré-processados para carga em uma banco de dados, resolvendo problemas de formatação.
2. Cada texto é tokenizado e gramaticalmente anotado com seu respectivo rótulo POS (*part-of-speech*).
3. Cada sentença é analisada para se detectar nomes próprios e conceitos concretos.
4. Todas as menções identificadas nos textos são ligadas às respectivas entidades da base de entidades.
5. Todas as menções e as respectivas entidades são compiladas e exibidas como resultado do processamento.

A abordagem MEV descrita no capítulo 2 apresenta muitas semelhanças com o roteiro acima. Na tabela 5.1 é apresentado o relacionamento entre as etapas das duas metodologias.

Como visto na tabela 5.1, as etapas do processo de aplicação do MEV assemelha-se com a abordagem tradicional para a identificação e ligação de entidades. Primeiramente, os textos e a base de entidades são pré-processados com o objetivo de resolver problemas de formatação e eliminar palavras que agregam pouco ao sentido do texto (*stopwords*), e estabelecer uma estrutura única para o tratamento das informações, tanto para os textos como para a base de entidades. O resultado desse processamento é a carga da base com as entidades extraídas da Wikipédia (no caso deste trabalho), e a formatação dos textos que serão analisados automaticamente.

O segundo passo consiste na tokenização do texto. Neste processo, cada palavra é analisada utilizando-se uma ferramenta de taggeamento POS que identifica a forma normal

Tabela 5.1: Etapas para a ligação de entidades e a relação com o processamento MEV.

Ligação de Entidades	**MEV**
1. Textos e base pré-processadas para eliminar problemas de formatação.	Tokenização.
2. Tokenização dos textos e taggeamento POS.	Normalização, Anotação.
3. Identificação de menções (entidades nomeadas).	Construção das matrizes.
4. Escolha das entidades para cada menção.	Ponderação, suavização e seleção.
5. Compilação do resultado.	Compilação do resultado.

da palavra e a classificação sintática da palavra no contexto em que ela está inserida. Trata-se de uma etapa fundamental onde a precisão da ferramenta de taggeamento POS influencia diretamente na qualidade dos resultados obtidos. Neste trabalho foi utilizado o TreeTagger [44], uma ferramenta desenvolvida por Helmunt Schmid na Universidade de Sttutgart.

Em seguida, os tokens são avaliados no sentido de identificar a existiência de nomes próprios ou conceitos concretos. Nessa seleção, são utilizadas as classes de palavras que mais ocorrem na composição de nomes próprios e de conceitos concretos.

A quarta etapa consiste na seleção das entidades para cada uma das menções identificadas nos textos (nomes próprios e conceitos concretos). Nessa seleção as entidades candidatas serão representadas como MEV e por meio de consultas envolvendo as menções identificadas nos textos as entidades mais adequadas para cada menção serão obtidas.

Por fim, o resultado contendo os pares "menção" e "entidade" ligadas será produzido, contendo todos os nomes próprios e conceitos concretos extraídos dos textos.

O modelo como um todo é descrito nas subseções que seguem.

5.1.1 Criação da Base de Entidades

A identificação de entidades em textos pode ser empregada em diferentes áreas, tais como detecção de tópicos, tradução de textos e extração de informação, dentre outras. Paralelo à identificação de entidades e em se tratando de grandes volumes de textos, a resolução de ambiguidades ocupa papel central tendo em vista as semelhanças de significado e as diferentes formas de apresentação das palavras contidas em um texto [8].

A desambiguação de entidades, uma das etapas do processo de ligação de entidades, dedica-se a estabelecer critérios que possibilitem a correta classificação das entidades identificadas em um texto. Por exemplo, a entidade "George W. Bush" (ex-presidente dos

Estados Unidos) pode ser representada de diferentes formas, tais como “George Bush” e “Bush”, sendo que a menção “Bush” analisada isoladamente pode referir-se a diferentes entidades: ao ex-presidente, propriamente dito; ao jogador de futebol americano Reggie Bush; ou ainda à banda de rock Bush.

Do exemplo anterior percebe-se que a efetividade na desambiguação de uma entidade depende da existência de diferentes opções de classificação/ligação e do sentido atribuído à expressão que contém tal entidade [3]. Dessa forma, a utilização de grandes bases heterogêneas de conhecimento pode incrementar a precisão da tarefa de desambiguar entidades por conter alterativas de classificação para assuntos distintos.

Nos últimos anos, devido à proliferação de comunidades de compartilhamento de conhecimento, diversas bases de conhecimento passaram a estar disponíveis. Tais bases contém conhecimento valioso sobre entidades, suas propriedades semânticas e o relacionamento entre diferentes entidades. O principal exemplo disso é a própria Wikipédia, que atualmente conta com mais de 4 milhões de artigos em Inglês e aproximadamente 815 mil artigos em Português, caracterizando-se como um dos mais importantes insumos para a atividade de relacionar entidades e bases de conhecimento [16].

Outra importante característica da Wikipédia, que reforça sua utilização como base de entidades para ligação de entidades, é a quantidade de links existentes em cada artigo. Os colaboradores da Wikipédia inserem em seus textos, de maneira manual, links dos termos mais importantes para outras páginas, proporcionando ao usuário uma maneira mais rápida de se acessar informações adicionais [1].

Neste trabalho, a Wikipédia foi adotada como principal fonte de entidades. Os conceitos concretos identificados nos textos são ligados às entidades representadas por páginas da Wikipédia. Para isso, foi necessário desenvolver uma ferramenta em Java para a importação das bases da Wikipédia, disponíveis de maneira *offline*, para o banco de dados MySQL utilizado pela ferramenta.

Atualmente, a Wikipédia disponibiliza sua base *offline* de diversas maneiras:

- Artigos, templates, informações de mídia e metadados.
- Artigos completos com histórico de edição.
- Artigos e respectivas revisões.
- Log de eventos para os artigos.
- Artigos recombinados.
- Artigos completos, na versão atual.
- Artigos e respectivos resumos.

Como será detalhado nas seções seguintes, cada artigo da Wikipédia representa uma entidade no modelo MEV. Os nomes próprios (NP) e conceitos concretos (CC) extraídos dos textos serão comparados com todas as entidades candidatas que comporem o MEV. Para que ocorra uma seleção correta, é fundamental que as entidades selecionadas possuam a maior quantidade de informação possível, já que o processo de desambiguação depende do grau de informação sobre o contexto em que o artigo está inserido. A eficácia da abordagem MEV está diretamente relacionada à quantidade de informações presentes para cada entidade.

Retomando o exemplo de "George W. Bush", suponha que se deseja analisar um texto sobre a banda "Bush" noticiando a entrada de um novo integrante chamado "George Willis", e que seja obtido da Wikipédia dois artigos que contenham o nome "Bush": "George_W_Bush", o político; e "Bush", a banda. Caso não seja obtida mais nenhuma informação sobre o assunto tratado por cada artigo, é coerente dizer que a ferramenta pode escolher o artigo "George_W_Bush" como o mais apropriado para o texto em questão, já que os dois nomes mais importantes para o texto estão contidos no título do artigo. Porém, somente o segundo artigo trata da banda "Bush" realmente, e essa informação é obtida por meio de uma análise mais profunda sobre o artigo.

Infelizmente não é viável utilizar as versões completas de todos os artigos da Wikipédia devido à complexidade computacional necessária para processar tanta informação. Ao invés de analisar todos os artigos por inteiro, optou-se por trabalhar com a versão *offline* da Wikipédia que contém um resumo do assunto abordado, retirado das primeiras palavras dos artigos. Apesar de não se ter todas as informações disponíveis sobre cada artigo, a base adotada fornece informações suficientes para se realizar a desambiguação na maioria dos casos.

As bases *offline* importadas da Wikipédia, contendo os resumos dos artigos, totalizaram 23 arquivos (bases disponíveis em novembro/2014) de aproximadamente 4,5 Gb. De cada arquivo foram armazenados o "nome", a "URL" (retirando-se o prefixo "wikipedia.org") e o "resumo". No total foram importados aproximadamente 4,7 milhões de artigos, que juntos somam mais de 50 milhões de palavras de conteúdo. Um exemplo do arquivo importado da Wikipédia contendo o título, o link e o resumo das entidades é mostrado na figura 5.1.

```
<doc>
<title>Wikipedia: ECS Electrochemistry Letters</title>
<url>http://en.wikipedia.org/wiki/ECS_Electrochemistry_Letters</url>
<abstract>ECS Electrochemistry Letters is a monthly peer-reviewed scientific journal
covering electrochemical science and technology. It was established in 2002 and is
published by the Electrochemical Society.</abstract>
</doc>
```

Figura 5.1: Exemplo de como as páginas estão descritas nos arquivos da Wikipédia.

Como será visto adiante, além da base da Wikipédia, foi importada também a base de Wikilinks fornecida pelo Wise 2013, para avaliação da ferramenta em comparação com o estado da arte (UnBWikilinks [1]). A base de Wikilinks contém uma lista de URLs da Wikipédia e as respectivas menções, que são referências para as páginas da Wikipédia extraídas de outras fontes da web. Nessa importação, as menções foram armazenadas da mesma forma que os resumos dos artigos contidos na base da Wikipédia, já que o princípio é o mesmo: utilizar as menções para seleção das entidades (desambiguação).

5.1.2 Extração de Nomes Próprios e Conceitos Concretos

A escolha dos NP e CC começa com o tratamento dos textos de entrada. O primeiro passo é a tokenização das sentenças e a análise das palavras por meio da ferramenta TreeTagger[1]. A análise *part-of-speech* (POS) é necessária uma vez que a maior parte dos conceitos concretos são formados por palavras que pertencem a um pequeno conjunto de classes POS, com base nos resultados oficiais fornecidos pela organização Wise 2013, como pode ser visto na tabela 5.2.

Tabela 5.2: Classes de palavras formadoras de nomes próprios e conceitos concretos.

Classes Gramaticais	**Qtde.**	**Freq. Absoluta**	**Freq. Relativa**
NP	13.780	40,08%	40,08%
NP NP	7.530	21,90%	61,99%
NN	2.137	6,22%	68,20%
NP NP NP	1.789	5,20%	73,41%
JJ NN	1.551	4,51%	77,92%
NN NN	1.396	4,06%	81,98%
NNS	822	2,39%	84,37%
NN NNS	785	2,28%	86,65%
JJ NNS	669	1,95%	88,60%
DT NP	389	1,13%	89,73%
NPS	353	1,03%	90,76%
JJ	351	1,02%	91,78%
NP NPS	244	0,71%	92,49%
NP NN	242	0,70%	93,19%
DT NP NP	220	0,64%	93,83%
DT NP NP NP	215	0,63%	94,46%
NN NP	125	0,36%	94,82%
JJ NP	122	0,35%	95,18%
NP NP NP NP	77	0,22%	95,40%
DT NN	66	0,19%	95,59%

[1]http://nlp.stanford.edu/software/tagger.shtml.

A tabela 5.2 lista as categorias de POS presentes nos NP e CC do resultado oficial do Wise 2013 mais comuns. A primeira coluna indica a combinação de diferentes categorias POS em uma mesma sentença; a segunda coluna indica o número de ocorrências de cada combinação da primeira coluna no resultado oficial; e as duas últimas colunas indicam a frequência destas ocorrências. A segunda linha da tabela, por exemplo, deve ser interpretada da seguinte forma: nomes próprios que consistem em apenas uma palavra, classificada como NP (nome próprio no singular), representam 40,08% de todas as ligações de entidades realizadas.

As categorias listadas na tabela 5.2 representam: adjetivos (JJ); substantivos no singular (NN); substantivos no plural (NNS); nomes próprios no singular (NP); nomes próprios no plural (NPS); e o determinante (DT). Pela análise da tabela é possível perceber que mais de 92% dos conceitos concretos e nomes próprios são formados de substantivos e adjetivos.

Com base nessa análise, neste trabalho as menções com probabilidade de formarem NP ou CC serão aquelas formadas apenas por palavras pertencentes às classes JJ, NN, NNS, NP e NPS. Dessa forma, para cada frase contida nos textos analisados, todas as expressões (combinações de palavras em ordem) possíveis são obtidas, mas são analisadas apenas aquelas que forem formadas exclusivamente pelas classes de palavras indicadas acima, conforme exemplo abaixo:

The/DT TreeTagger/NP is/VBZ easy/JJ to/TO use/VB.

Da frase acima podemos apontar as seguintes combinações:

1. The TreeTagger is easy to use
2. The TreeTagger is easy to
3. TreeTagger is easy to use
4. The TreeTagger is easy
5. TreeTagger is easy to
6. is easy to use
7. The TreeTagger is
8. TreeTagger is easy
9. is easy to
10. easy to use
11. The TreeTagger

12. TreeTagger is
13. is easy
14. easy to
15. to use
16. The
17. TreeTagger
18. is
19. easy
20. to
21. use

No exemplo acima, a única combinação formada apenas pelas classes de palavras indicadas anteriormente é *TreeTagger*, formada por uma palavra NP. As demais contém palavras de outras classes de POS que em princípio não formam NP ou CC, tais como verbos, advérbios, artigos, etc, e por isso não são consideradas como candidatas a NP ou CC. A combinação *TreeTagger* então será analisada por meio do MEV a fim de se obter a entidade mais adequada.

Após a identificação de todos os NP e CC presentes no texto, inicia-se a etapa de identificação das entidades candidatas e a escolha da entidade mais adequada para a LE, realizada no passo seguinte. A lista de entidades candidatas é formada por todas as entidades que contém, seja no nome ou no resumo, as palavras contidas no NP ou CC analisado. No exemplo acima, todas as entidades (páginas da Wikipédia) que contém a palavra *TreeTagger* em seu nome (link) ou citada em seu resumo serão consideradas como candidatas para a LE.

5.1.3 Ligação dos Nomes Próprios e Conceitos Concretos com Entidades Wiki

A idéia da abordagem MEV é de representar cada documento em uma coleção como um ponto no espaço multidimensional (um vetor em um espaço vetorial). Pontos no espaço vizinhos são semanticamente similares e pontos distantes são semanticamente diferentes. As consultas também são representadas como um ponto no mesmo espaço (a consulta é um tipo de pseudo-documento). Os documentos são classificados em ordem crescente de acordo com a distância a partir da consulta [54].

Os Modelos de Espaço Vetorial extraem automaticamente o conhecimento de conjuntos de textos, exigindo menos trabalho do que outras abordagens semânticas, como ontologias. Além disso, MEV pode ser aplicado em tarefas que precisem do cálculo da semelhança entre palavras, frases e documentos. MEV é especialmente interessante por causa da sua relação com a *hipótese distributiva*: palavras que ocorrem em contextos semelhantes tendem a ter significados semelhantes [54].

A abordagem de MEV foi adotada neste trabalho para a seleção das entidades mais adequadas para cada conceito concreto extraídos do texto. Para cada conceito concreto, uma matriz é construída para representar as entidades candidatas e o conceito concreto que se deseja ligar. A representação MEV adotada segue a metodologia descrita no capítulo 2, com pequenas adaptações.

Neste trabalho a seguinte notação é adotada: seja $\mathbf{X}$ uma matriz *entidade-termo*. Suponha que a coleção contenha n entidades e m termos únicos. A matriz $\mathbf{X}$ terá, então, m linhas (cada linha representa um termo presente em uma entidade) e n colunas (cada coluna representa uma entidade). Suponha, ainda, que w_i representa o i-ésimo termo na base de entidades e d_j a j-ésima entidade na coleção. A i-ésima linha em $\mathbf{X}$ é o vetor linha $\mathbf{x}_{i:}$ e a j-ésima coluna em $\mathbf{X}$ é o vetor coluna $\mathbf{x}_{:j}$. O vetor linha $\mathbf{x}_{i:}$ contém n elementos, um elemento para cada entidade, e o vetor coluna $\mathbf{x}_{:j}$ contém m elementos, um elemento para cada termo. Suponha que $\mathbf{X}$ é uma simples matriz de frequências. O elemento x_{ij} em $\mathbf{X}$ é a frequência do i-ésimo termo w_i na j-ésima entidade d_j.

O vetor $\mathbf{x}_{:j}$ é considerado uma representação da entidade j. Ele nos diz com que frequência os termos aparecem na entidade, mas a ordem sequencial dos termos é desconsiderada. O vetor não tem como objetivo capturar a estrutura de frases, sentenças, parágrafos e capítulos da entidade ou do texto analisado. No entanto, apesar da simplicidade, os motores de busca funcionam surpreendentemente bem com essa representação, demonstrando que vetores são capazes de captar importantes aspectos semânticos [54].

A tabela 5.3 apresenta um exemplo de uma matriz em que cada linha de frequências representa um termo e cada coluna representa uma entidade, ou seja, o termo *Antony* ocorre 157 vezes na entidade *Antony Cleopatra*, e outras, enquanto que o termo *Calpurnia* ocorre 10 vezes na entidade *Julius Caesar*, e somente nesta entidade.

Após da tokenização, normalização (opcional) e anotação, o próximo passo consiste em gerar a matriz de frequências descrita acima. Em seguida, os pesos dos elementos da matriz são ajustados pois as palavras de alta frequência são menos relevantes do que as palavras raras. Finalmente, a semelhança entre os vetores é calculada (existem diferentes maneiras de se realizar esta etapa).

O objetivo da ponderação dos elementos da matriz, como visto no capítulo 2, é aplicar um peso maior para os eventos mais importantes e menos peso para eventos esperados.

Tabela 5.3: Um exemplo de uma matriz de frequências.

	Antony Cleopatra	**Julius Caesar**	**The Tempest**
Antony	157	73	4
Brutus	4	157	0
Caesar	232	227	0
Calpurnia	0	10	0
Cleopatra	57	0	0
mercy	2	0	3
worser	2	0	1

A hipótese é que eventos raros que são compartilhados por dois vetores indicam maior semelhança entre os vetores. Em teoria da informação, eventos raros possuem mais conteúdo do que eventos esperados [45]. A forma mais popular de formalizar esta ideia para matrizes *termo-documento* é a família TF-IDF (frequência do termo *versus* inverso da frequência no documento) de funções de ponderação [19], descrita no capítulo 2. Um elemento recebe um peso elevado quando o termo correspondente é frequente no documento correspondente (ou seja, TF é alta), mas o termo é raro em outros documentos do conjunto (ou seja, DF é baixa e, assim, IDF é alta). As funções de ponderação da família TF-IDF podem produzir melhoras significativas em tarefas de recuperação de informação quando comparadas com a frequência bruta [42]. O raciocínio é o mesmo para matrizes *entidade-termo*.

Na representação proposta, por meio de matrizes *entidade-termo*, as componentes TF-IDF dos elementos das matrizes, ou seja, tf e idf_t, respectivamente, bem como os valores ponderados dos elementos, são obtidos por meio das equações 2.1, 2.2 e 2.3 descritas no capítulo 2.

A tabela 5.4 mostra a mesma matriz da tabela anterior, após o cálculo da TF-IDF. O resultado já considera a normalização da matriz:

Tabela 5.4: Um exemplo de uma matriz de frequências após o cálculo da TF-IDF.

	Antony Cleopatra	**Julius Caesar**	**The Tempest**
Antony	0,000	0,000	0,000
Brutus	0,187	0,448	0,000
Caesar	0,394	0,471	0,000
Calpurnia	0,000	0,760	0,000
Cleopatra	0,874	0,000	0,000
mercy	0,152	0,000	0,828
worser	0,152	0,000	0,561

A similaridade entre os vetores da representação MEV adotada é calculada através da função cosseno, obtida por meio da equação 2.7 apresentada, também, no capítulo 2. Na abordagem proposta, o cálculo do cosseno é realizado entre cada vetor-entidade e o

vetor-consulta que representa o NP ou CC analisado, formado pelos termos presentes no texto, com o objetivo de representar o contexto no qual o NP ou CC está inserido.

O uso de MEV revela-se bastante adequado para o modelo proposto. No entanto, a falta de um conjunto homogêneo de dados pode dar origem a resultados insatisfatórios. Um exemplo é a existência de um banco de dados formado por artigos da Wikipédia com conteúdos de diferentes tamanhos e/ou qualidades. Uma página de um candidato forte, mas contendo apenas algumas palavras de conteúdo, pode ser descartada em detrimento de outra página menos relacionada ao assunto, mas que contenha vários parágrafos com um maior número de citações sobre o conteúdo pesquisado, já que o método é baseado na frequência dos termos.

Para superar o problema da falta de homogeneidade, propõe-se o uso da função de ajuste a seguir:

$$\phi(n_i, e_j) = \alpha A + \beta B \tag{5.1}$$

Onde, n_i representa as entidades candidatas para cada conceito e_j extraído do texto, A é o resultado obtido no cálculo do cosseno entre as entidades candidatas e o conceito concreto avaliado no MEV, e B indica a correlação do conceito concreto com as palavras contidas nos nomes das entidades e com o tamanho destes mesmos nomes (quantidade de palavras). Uma vez que não se pode garantir a qualidade ou a consistência do conteúdo dos artigos da Wikipédia, sugere-se avaliar a correlação do conceito concreto com o nome dado aos artigos candidatos. A escolha dos nomes dos artigos é uma tarefa realizada cuidadosamente já que, em poucas palavras, os autores procuram definir da melhor maneira a questão a ser abordada no artigo.

O *alfa* e o *beta* são constantes/pesos atribuídos aos dois parâmetros calculados (cosseno/MEV e correlação com títulos) o que permite equilibrar a sensibilidade dos cálculos. Nas avaliações apresentadas no capítulo 6 os valores de *alfa* e *beta* foram obtidos a partir de testes realizados com um grupo de treinamento contendo textos que não fizeram parte das avaliações. Nesses testes foram utilizadas diferentes combinações de valores para os dois parâmetros, e a combinação que apresentou maiores índices de *recall* para o conjunto de treinamento foi adotada nas avaliações do capítulo 6.

A correlação com os títulos dos artigos considera a extensão do nome do artigo e a quantidade de correspondências encontradas. Por exemplo, para um conceito concreto formado por duas palavras, a ordem de importância (ordem decrescente) das entidades candidatas é: 1) entidade que tem apenas as mesmas duas palavras na URL (ou título); 2) entidades que contêm várias palavras na URL, incluindo estas duas palavras (URL mais longa); 3) entidades que contêm apenas uma dessas duas palavras na URL; e 4) entidades

que não contenham qualquer uma dessas duas palavras na URL. O fator B capta essa correlação.

5.2 Arquitetura da Solução

A figura 5.2 mostra a arquitetura da solução proposta, indicando os passos realizados e as transições de fluxo que ocorrem entre eles.

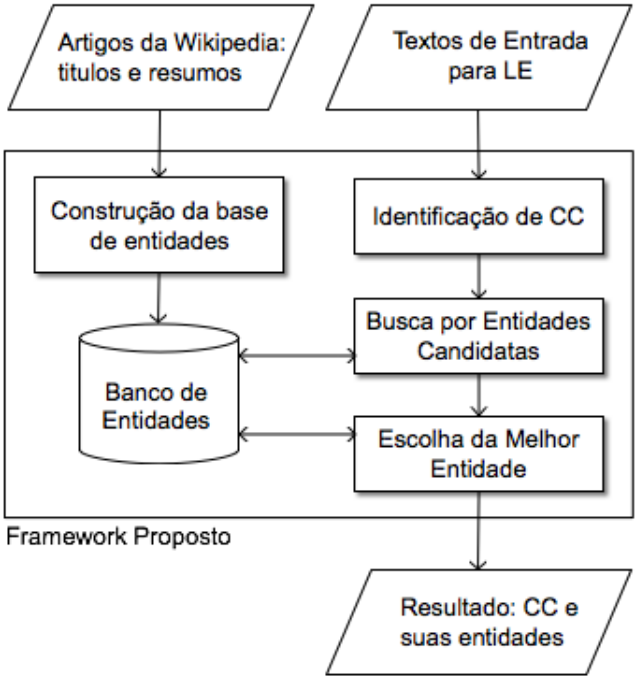

Figura 5.2: *Overview* da arquitetura proposta.

Na figura 5.2 estão representados, de fora do núcleo do framework, as entradas dos arquivos referentes às bases da Wikipédia (passo 1) e dos textos a serem analisados (passo 2), e a saída do processo como um todo que consiste nos pares de NP e CC e respectivas entidades (passo 3).

Já no núcleo do framework estão representadas as atividades de criação da base de entidades e a base propriamente dita, cujo responsável é o passo 1, na figura à esquerda. Na parte da direita estão as atividades dos demais passos: identificação de conceitos concretos (passo 2); e pesquisa por entidades candidatas e seleção das melhores entidades (ambos do passo 3).

5.3 Resumo do Capítulo

Neste capítulo foi apresentada a metodologia proposta para LE envolvendo conceitos concretos e entidades Wiki. A metodologia apresentada executa três passos fundamentais, e segue de certa forma o framework comum adotado pelos trabalhos que investigam a tarefa de LE: detecção das sentenças ligáveis nos textos avaliados; classificação e seleção das respectivas entidades; e desambiguação das entidades selecionadas para cada menção.

Os passos em questão foram agrupados nas etapas "extração de nomes próprios e conceitos concretos" e "ligação dos nomes próprios e conceitos concretos com as respectivas entidades Wiki" que são precedidas pela etapa de "criação da base de entidades", fundamental para a realização da LE. Cada uma dessas etapas foi explorada em detalhes no capítulo, com as devidas referências aos conteúdos citados nos capítulos anteriores, principalmente no que diz respeito à fundamentação teórica envolvendo o uso de Modelos de Espaço Vetorial.

No próximo capítulo o modelo proposto será colocado em prática através da execução da ferramenta UnBWiki VSM que implementou os conceitos apresentados neste capítulo.

Capítulo 6

Estudo de Caso

Neste capítulo são apresentados dois estudos de caso para validação e análise do comportamento do modelo proposto. O primeiro estudo de caso compreende a utilização da ferramenta com dados já tratados por outros trabalhos que investigaram a tarefa de ligação de entidades envolvendo conceitos concretos, possibilitando uma comparação direta dos resultados obtidos pela ferramenta UnBWiki VSM com o estado da arte. O segundo estudo de caso, realizado com dados extraídos da web, tem como objetivo avaliar a utilização da ferramenta em uma situação real para verificação da aplicabilidade dos resultados obtidos.

6.1 Estudo de Caso 1: Wise

O primeiro estudo de caso tem como objetivo avaliar o desempenho da ferramenta UnBWiki VSM em comparação com a ferramenta UnBWikilinks. A ferramenta UnBWikilinks, cuja metodologia foi detalhada no capítulo 4, foi apresentada no evento Wise 2013 como solução para o desafio de ligação de entidades (LE) envolvendo conceitos concretos, e obteve resultados satisfatórios conforme resultado oficial divulgado pela organização do evento.

Nas próximas subseções os dados serão descritos e os resultados obtidos pela ferramenta UnBWiki VSM serão comparados com aqueles obtidos pela ferramenta UnBWikilinks.

6.1.1 Descrição dos Dados

Neste estudo de caso foi utilizado o conjunto de textos fornecidos pelo Wise 2013. O referido conjunto é composto por 8.800 textos sobre os mais diversos assuntos, e os participantes do evento receberam o desafio de processar os textos com o objetivo de

identificar pares de conceitos concretos e as respectivas entidades. As entidades foram recuperadas do banco de dados contendo wikilinks.

A versão original da presente ferramenta trabalha com um banco de dados formado por artigos da Wikipédia e os respectivos resumos. A escolha do resumo do artigo como característica relevante na análise da similaridade se deve a dois fatores principais: a) o resumo tende a ter a informações-chave sobre o assunto abordado no artigo e, consequentemente, reduz o risco de se tratar textos muito longos que possam mudar o foco do artigo; e, b) a utilização de artigos completos aumenta significativamente a complexidade do processamento devido ao imenso número de palavras que precisam ser analisadas pela ferramenta.

A ferramenta UnBWikilinks foi avaliada utilizando um banco de dados formado por wikilinks, que consiste em uma lista de URLs de artigos da Wikipédia e suas referências obtidas a partir de diferentes páginas da web. Para efeitos de avaliação da metodologia proposta, o banco de dados original formado com artigos da Wikipédia foi substituído pelo banco de dados Wikilinks fornecido pela organização do Wise 2013.

O banco de dados de wikilinks possui mais de 2,8 milhões de entidades sobre os mais variados asuntos, com mais de 18 milhões de palavras, sendo mais de 1 milhão de palavras distintas. Dos 1.500 textos analisados manualmente pelos organizadores do Wise 2013, 5% (ou seja, 75 textos) foram processados com a ferramenta UnbWiki VSM neste estudo de caso.

O uso do banco de dados de wikilinks possibilita a correlação direta dos conceitos concretos com as respectivas entidades, devido ao fato de que as citações recolhidas a partir de outras páginas da web são, por si só, um forte indicativo de como esses conceitos concretos são tradicionalmente utilizados, ou seja, em quais contextos tais conceitos concretos são aplicados. Por outro lado, a Wikipédia fornece um banco de dados mais rico, com uma maior quantidade de conteúdo que é continuamente atualizado.

Para realizar esta comparação, foi utilizado o mesmo banco de dados empregado na abordagem UnBWikilinks. Essa adaptação é justificável por se tratar de um trabalho pioneiro na ligação de entidades envolvendo conceitos concretos. Assim, para avaliar a ferramenta, base de dados da Wikipédia, que é a base original da metodologia, foi substituída pelo banco de dados contendo wikilinks, com um pequeno ajuste que não afeta a utilização da metodologia: os resumos dos artigos da Wikipédia da abordagem original foram substituídos pelas menções aos artigos contidas na base Wikilinks. Dessa forma, a metodologia ora proposta pode ser comparada com outra abordagem (UnBWikilinks) em um cenário contendo a mesma base de entidades e os mesmos textos passados como entrada, obtendo-se resultados diretamente comparáveis.

6.1.2 Simulação e Resultados

A organização do Wise 2013 forneceu mais de 8.800 textos da Internet, sobre vários assuntos, para que os participantes pudessem processar nos testes de suas metodologias. Desse conjunto, 1.500 textos foram analisados manualmente pelos organizadores para formar um conjunto de referência, um gabarito oficial a ser comparado com os resultados obtidos pelos participantes.

A avaliação do Wise 2013 considerou três indicadores: a) *recall* na obtenção de nomes próprios (NP); b) *recall* na obtenção de nomes próprios e conceitos concretos (CC) combinados; e, c) *precisão* na obtenção de nomes próprios e conceitos concretos combinados. Os três melhores resultados verificados pela organização do Wise 2013 são mostrados na tabela 4.1.

A tabela 4.1 mostra que a ferramenta UnBWikilinks obteve *recall* igual a 40,1%, próximo ao *recall* obtido pelas equipes 299 (47,5%) e 306 (44,1%), como pode ser verificado na segunda coluna da tabela. A terceira coluna apresenta o *recall* obtido por cada uma das ferramentas na ligação de nomes próprios, que não foi levado em consideração uma vez que o foco dos organizadores era a LE de conceitos concretos. A última coluna da tabela mostra que a *precisão* da ferramenta UnBWikilinks na LE de NP e de CC combinados (42,5%) foi consideravelmente superior à obtida pelos demais participantes: 52,9% superior à obtida pela equipe 306, segundo melhor resultado; e 196,6% superior à obtida pela equipe 299, terceira colocada.

A análise do resultado oficial mostrou que, provavelmente em função do gabarito ter sido produzido manualmente pela organização do Wise 2013, tanto a *precisão* como o *recall* obtidos pelos participantes poderiam estar distorcidos (a *precisão* com maior distorção). Uma das distorções pode ocorrer em função de situações como a ocorrência de um mesmo conceito concreto em diferentes trechos do texto, porém foram considerados apenas algumas ocorrências no gabarito oficial. Exemplificando, vamos supor que em um determinado texto exista apenas o conceito concreto *idioma estrangeiro*, e que ele ocorra em três partes do texto, mas que o gabarito oficial tenha considerado somente duas ocorrências. Uma equipe que tenha identificado os três conceitos concretos corretamente, e somente eles, deveria receber 100,0% como valores de *precisão* e *recall*, porém pelo gabarito oficial receberia 66,7% e 100,0%, respectivamente. Uma segunda equipe que tenha identificado duas ocorrências do conceito concreto, e somente as duas, receberia os 100,0% nos dois indicadores de maneira indevida, quando deveria receber 66,7% de *recall*.

Um outro exemplo dessa distorção está relacionado à existência de vários conceitos concretos de uma mesma natureza. Exemplificando, vamos supor que em um determinado texto existam os conceitos concretos *soccer*, *basketball* e *boxing*, mas que o gabarito oficial

tenha considerado apenas o CC *boxing*. Uma equipe que tenha recuperado os três conceitos concretos receberia como *precisão* o valor de 33,3% quando o correto seria 100,0%.

As distorções acima, além de outras identificadas, podem ser atribuídas ao fato de que a análise de conceitos concretos carrega certa ambiguidade. O que pode ser um conceito concreto para um analista não será necessariamente um conceito concreto para outro analista. Essa característica dificulta o alcance de um resultado 100,0% unânime.

Visando diminuir o impacto das distorções em questão, na avaliação deste estudo de caso foi considerado apenas o indicador *recall* uma vez que não é possível garantir que apenas os conceitos concretos contidos no resultado oficial estão corretos. Pelos exemplos acima é possível verificar que a *precisão* é o indicador que sofre mais impacto com essas distorções. Além disso, o indicador *recall* de NP não foi levado em conta já que o foco da ferramenta é LE envolvendo CC.

Neste estudo de caso 75 textos dos 1.500 documentos disponibilizados pelo Wise 2013 foram selecionados aleatoriamente e submetidos ao processamento das duas ferramentas, UnBWikilinks e UnBWiki VSM. Nesse conjunto de 75 textos foram identificados, segundo o gabarito do Wise 2013, 930 NP e CC combinados. Os resultados obtidos pelas ferramentas foram analisados de três formas diferentes.

Na primeira análise os textos foram agrupados de acordo com a quantidade de NP e CC presentes, da seguinte forma: um grupo contendo textos com até 5 NP e CC combinados; um segundo grupo contendo entre 6 e 10 NP e CC combinados; um terceiro grupo contendo entre 11 e 15 NP e CC combinados; um quarto grupo contendo entre 16 e 20 NP e CC combinados; e um quinto grupo contendo 21 ou mais NP e CC combinados.

Tabela 6.1: *Recall* UnBWiki VSM *versus* UnBWikilinks - textos agrupados pela quantidade de NP e CC.

Qtde. NP / CC	**Gabarito**	**Wikilinks**	$R_{wikilinks}$	**Wiki VSM**	R_{VSM}
0 a 5	57	32	40,4%	45	47,4%
6 a 10	146	78	37,7%	118	48,6%
11 a 15	197	101	33,0%	156	43,1%
16 a 20	285	131	30,2%	218	42,1%
21 ou mais	245	125	35,5%	201	49,4%
Total	**930**	**467**	**34,0%**	**738**	**45,6%**

Os resultados obtidos pelas duas ferramentas estão descritos na tabela 6.1. A primeira coluna identifica cada um dos cinco grupos, de acordo com a quantidade de NP e CC combinados presentes nos textos. A segunda coluna contém a quantidade de NP e CC presentes nos textos de cada grupo, conforme o gabarito oficial do Wise 2013. A terceira coluna indica a quantidade de NP e CC identificados pela ferramenta UnBWikilinks, e o *recall* calculado está descrito na coluna $R_{wikilinks}$. Por fim, a quinta coluna indica a

quantidade de NP e CC identificados pela ferramenta UnBWiki VSM, e o *recall* calculado está descrito na coluna R_{VSM}.

Como mostrado na tabela 6.1, os resultados obtidos pela ferramenta UnBWiki VSM foram superiores aos resultados obtidos pela ferramenta UnBWikilinks em todos os grupos. Os resultados demonstram, ainda, que à medida em que a quantidade de NP e CC combinados aumenta, a performance da ferramenta UnBWiki VSM é ainda melhor do que a da ferramenta UnBWikilinks. No grupo de textos com até 5 NP e CC combinados a metodologia deste trabalho obteve *recall* 17,3% superior ao da ferramenta UnBWikilinks. Já no grupo de textos com 21 ou mais NP e CC combinados, o *recall* obtido pela ferramenta UnBWiki VSM foi superior em 39,2%.

Na segunda análise os textos foram agrupados de acordo com o desempenho comparativo entre as duas ferramentas: o primeiro grupo contém os textos cujo *recall* obtido pela ferramenta UnBWiki VSM foi pior do que o *recall* obtido pela ferramenta UnBWikilinks; o segundo grupo contém os textos cujo *recall* obtido pelas duas ferramentas foi semelhante; e o terceiro grupo contém os textos cujo *recall* obtido pela ferramenta UnBWiki VSM foi superior ao *recall* obtido pela ferramenta UnBWikilinks.

Tabela 6.2: *Recall* UnBWiki VSM *versus* UnBWikilinks - textos agrupados pelo desempenho comparativo entre as duas ferramentas.

Desempenho	**Gabarito**	$\boldsymbol{R_{wikilinks}}$	$\boldsymbol{R_{VSM}}$	**Variação**
Pior	52	44,2%	28,8%	-34,8%
Igual	235	43,4%	43,4%	0,0%
Melhor	643	29,7%	47,7%	60,7%
Total	**930**	**34,0%**	**45,6%**	**34,2%**

Os resultados obtidos pelas duas ferramentas estão descritos na tabela 6.2. A primeira coluna identifica cada um dos três grupos, de acordo com o desempenho comparativo entre as duas ferramentas. A segunda coluna contém a quantidade de NP e CC pertencentes a cada um dos três grupos. A terceira coluna indica o *recall* obtido pela ferramenta UnBWikilinks na análise dos textos de cada um dos grupos. A quarta coluna indica o *recall* obtido pela ferramenta UnBWiki VSM na análise dos textos de cada um dos grupos. Por fim, a coluna variação apresenta a variação percentual entre o *recall* das duas ferramentas.

Como mostrado na tabela 6.2, os resultados obtidos pela ferramenta UnBWiki VSM foram superiores aos resultados obtidos pela ferramenta UnBWikilinks na maioria dos casos. No grupo de textos em que a metodologia proposta neste trabalho foi melhor estão concentradas 69,1% dos NP e CC existentes no conjunto de textos de teste, com *recall* superior ao da ferramenta UnBWikilinks em 60,7%. Somente em 6 textos dos 75 analisados, ou seja, 8% do total, o desempenho da ferramenta UnBWiki VSM obteve

recall inferior. Por outro lado, em 56% dos textos analisados o método proposto obteve desempenho melhor do que o da ferramenta UnBWikilinks.

Na terceira análise os textos foram agrupados considerando o assunto abordado (contexto). Eles foram divididos em 7 categorias principais: *Saúde*, *Política*, *Notícia*, *Economia*, *Educação*, *Esporte* e um último grupo contendo textos de assuntos variados, denomiado *Outros*. A distribuição da quantidade de NP e CC contidos no gabarito oficial entre os grupos pode ser verificada na tabela 6.3.

Tabela 6.3: *Recall* UnBWiki VSM *versus* UnBWikilinks - textos agrupados pelo assunto abordado (contexto dos textos).

Assunto	**Gabarito**	$R_{wikilinks}$	R_{VSM}	**Variação**
Educação	62	16,1%	27,4%	70,0%
Economia	153	30,7%	41,8%	36,2%
Política	272	34,9%	47,1%	34,7%
Notícia	114	34,2%	45,6%	33,3%
Saúde	60	58,3%	66,7%	14,3%
Esporte	10	20,0%	20,0%	0,0%
Outros	259	34,0%	46,7%	37,4%
Total	**930**	**34,0%**	**45,6%**	**34,2%**

A primeira coluna da tabela 6.3 identifica cada um dos 7 grupos, de acordo com o assunto abordado pelos textos pertencentes a cada grupo. A segunda coluna contém a quantidade de NP e CC pertencentes a cada um dos grupos. A terceira coluna indica o *recall* obtido pela ferramenta UnBWikilinks na análise dos textos de cada um dos grupos. A quarta coluna indica o *recall* obtido pela ferramenta UnBWiki VSM na análise dos textos de cada um dos grupos. Por fim, a coluna variação apresenta a variação percentual entre o *recall* das duas ferramentas.

Como mostrado na tabela 6.3, os resultados obtidos pela ferramenta UnBWiki VSM só não foram superiores aos da ferramenta UnBWikilinks em um dos grupos, *Esporte*, onde o desempenho foi o mesmo. Nos demais grupos o desempenho da metodologia ora proposta foi superior, com destaque para os textos sobre *Educação* onde o *recall* obtido pela ferramenta UnBWiki VSM foi 70% superior ao da ferramenta UnBWikilinks. Analisando a ferramenta UnBWiki VSM individualmente, os textos sobre *Saúde* e *Política* merecem destaque, com *recall* de 66,7% e 47,1%, respectivamente.

Os resultados apresentados nas tabelas 6.1, 6.2 e 6.3 demonstram que a nova abordagem para o problema de LE envolvendo conceitos concretos e entidades Wiki apresenta resultados satisfatórios quando comparados com o estado da arte.

6.2 Estudo de Caso 2: Royaltree

O segundo estudo de caso tem como objetivo avaliar o desempenho da ferramenta UnBWiki VSM quando aplicada a um caso real. Nesse estudo de caso, foram selecionados textos sobre a família real britântica disponíveis na web, com o objetivo de avaliar se a ferramenta consegue identificar, nos textos analisados, nomes próprios e conceitos concretos para páginas da Wikipédia.

Figura 6.1: Exemplo de LE com Royaltree - conceito concreto *rei anglo-saxônico*.

Os dados utilizados no presente estudo de caso e as análises efetuadas são detalhadas nas subseções que seguem.

6.2.1 Descrição dos Dados

O site *British Royal Family History*[1] disponibiliza uma série de informações sobre a história da família real britânica. Na área dedicada à árvore genealógica (*Royaltree*), é possível navegar entre diferentes famílias, acessando informações sobre reis e seus descendentes, revelando fatos importantes sobre diversos momentos da história.

As informações presentes na página *Royaltree* são exibidas como forma de texto puro, não havendo links entre fatos históricos correlacionados, e desta forma se apresenta como

[1] http://www.britroyals.com/royaltree.htm

um estudo de caso adequado para a realização de ligação de entidades com páginas Wiki. Um exemplo da LE envolvendo *Royaltree* pode ser consultado na figura 6.1

O estudo de caso foi realizado com um conjunto de 5 textos da *Royaltree* escolhidos aleatoriamente. Desses textos, foram considerados apenas as primeiras frases, que trazem informações gerais, tendo sido descartadas as *tags* apresentadas no início das páginas, tais como nome, filiação, período, etc. Os textos selecionados foram: rei "Alfred, o Grande"; rei "Henrique VII"; rei "William IV"; rainha "Elizabeth II"; e "Tudors Mary Queen of Scots".

Como não existe gabarito oficial disponível, antes da realização dos testes os textos foram analisados manualmente, da seguinte forma: sentenças cujo analista julgou tratar-se de conceito concreto ou nome próprio foram destacadas do texto, e a sentença foi pesquisada na Wikipédia por meio de desambiguação manual. Essa análise manual formou o gabarito utilizado como referência para a avaliação da ferramenta.

Uma vez que o gabarito citado anteriormente pode provocar, também, as mesmas distorções verificadas quando da análise do gabarito oficial disponibilizado pelos organizadores do Wise 2013, na avaliação deste estudo de caso foi considerado apenas o indicador *recall* uma vez que não é possível garantir que apenas conceitos concretos presentes no gabarito, selecionados manualmente pelo analista, estão corretos.

6.2.2 Simulação e Resultados

Os 5 textos obtidos no site *Royaltree* foram processados pela ferramenta UnBWiki VSM da seguinte forma: as primeiras frases de cada texto foram copiadas da página web e salvas em arquivo contendo texto puro; e cada texto, em seguida, foi submetido ao processamento da ferramenta UnBWiki VSM, da mesma maneira do processamento realizado no estudo de caso envolvendo os textos do Wise 2013. Os fragmentos analisados pela ferramenta são listados na tabela 6.4.

A tabela 6.4 lista os membros da família real sobre os quais os textos fazem referência, na primeira coluna, e os fragmentos dos textos processados na segunda coluna. Cada linha da tabela, da linha 2 à linha 6, contém os dados de um texto e os respectivos NP e CC. As sentenças na segunda coluna destacados em *itálico* representam os NP e CC classificados indevidamente, segundo o gabarito manual. Já as sentenças destacadas em **negrito**, também na segunda coluna, representam os NP e CC classificados corretamente segundo o gabarito manual, ou seja, as sentenças foram descatadas, do jeito que estão, pelo analista que classificou os textos de maneira manual.

Na análise do texto *King Alfred, The Great* a ferramenta identificou corretamente 7 NP e CC e de maneira indevida 1 NP e CC (*Danish invasion*), conforme gabarito gerado manualmente. Os pares de "NP/CC : entidade" selecionados pela ferramenta durante a realização da LE são:

Tabela 6.4: Resultado da análise dos textos extraídos de *Royaltree*.

Texto *Royaltree*	**Fragmento Analisado**
King Alfred The Great	**Anglo-Saxon king** 871-899 who defended **England** against *Danish invasion* and founded the first **English navy**. He succeeded his brother **Aethelred** to the throne of **Wessex** in 871, and a new **legal code** came into force during his reign. He encouraged the translation of scholarly works from Latin (some he translated himself), and promoted the development of the **Anglo-Saxon Chronicle**.
King Henry VII	*Henry* 871-899 was the son of **Edmund Tudor, Earl** of Richmond, who died before *Henry* was born, and **Margaret Beaufort**, a descendant of **Eduard III** through John of **Gaunt, Duke** of Lancaster.
King William IV	*William* was the third son of **George III** and not expected to become king. He was sent off to join the **Royal Navy** at 13 years old, and saw service at the *Battle of St Vincent* against the Spanish in 1780 and in *New York* during the **American War** of **Independence**.
Queen Elizabeth II	**Princess Elizabeth Alexandra Mary** was born in **London** on 21 April 1926; she was educated privately, and assumed official duties at 16. During **World War II** she served in the **Auxiliary Territorial Service**, and by an amendment to the **Regency Act** she became a **state counsellor** on her 18th birthday.
Tudors Mary Queen of Scots	*Mary Queen of Scots* daughter of **James V** of *Scotland* was born at **Linlithgow Palace**, *West Lothian, Scotland*, on 8 December 1542, and became **Queen** of Scots when she was six days old.

- Anglo-Saxon king : Anglo-Saxon_monarchs
- England : Winchester,_England
- English navy : English_Navy
- Aethelred : Aethelred
- Wessex : Cerdic_of_Wessex
- legal code : Legal_code
- Anglo-Saxon Chronicle : Anglo-Saxon_chronicle

Já análise do texto *Henry VII* a ferramenta identificou corretamente 4 NP e CC e de maneira indevida 2 NP e CC (*Henry*, duas ocorrências), conforme gabarito gerado

manualmente. Os pares de "NP/CC : entidade" selecionados pela ferramenta durante a realização da LE são:

- Edmund Tudor, Earl : Edmund_Tudor,_1st_Earl_of_Richmond
- Margaret Beaufort : Margaret_Beaufort
- Eduard III : Eduard_III
- Gaunt, Duke : John_of_Gaunt,_Duke_of_Lancaster

Com relação ao texto *William IV* a ferramenta identificou corretamente 4 NP e CC e de maneira indevida 3 NP e CC (*William*, *Battle of St Vicent* e *New York*), conforme gabarito gerado manualmente. Os pares de "NP/CC : entidade" selecionados pela ferramenta durante a realização da LE são:

- George III : George_III
- Royal Navy : Royal_Navy
- American War : American_war_of_Independence
- Independence : war_of_independence

Na análise do texto *Queen Elizabeth II* a ferramenta encontrou os mesmos 6 NP e CC existentes no gabarito gerado manualmente. Os pares de "NP/CC : entidade" selecionados pela ferramenta durante a realização da LE são:

- Princess Elizabeth Alexandra Mary : Elizabet_II
- London : London
- World War II : World_War_II
- Auxiliary Territorial Service : Auxiliary_Territorial_Service
- Regency Act : Regency_Act
- state counsellor : state_counsellor

Por fim, na análise do *Tudors Mary Queen of Scots* a ferramenta identificou corretamente 3 NP e CC e de maneira indevida 3 NP e CC (*Mary Queen of Scots*, *Scotland* e *West Lothian, Scotland*), conforme gabarito gerado manualmente. Os pares de "NP/CC : entidade" selecionados pela ferramenta durante a realização da LE são:

- James V : James_V

- Linlithgow Palace : Linlithgow_Palace
- Queen : The_Queen_of_Scots

O resultado da análise dos textos acima estão compilados na tabela 6.5. A primeira coluna contém o nome do membro da família real aos quais cada um dos 5 textos faz referência. A coluna 2 contém a quantidade de NP e CC gabaritados manualmente pelo analista. Na coluna 3 são informadas as quantidades de NP e CC identificados corretamente pela metodologia ora proposta. Por fim, na coluna 4, são informados o *recall* para cada um dos 5 textos, bem como o *recall* total para o conjunto de textos.

Tabela 6.5: *Recall* UnBWiki VSM na análise dos textos da família real britânica.

Texto	**Gabarito**	**UnBWiki VSM**	R_{VSM}
King Alfred the Great	8	7	87,5%
King Henry VII	6	4	66,7%
King Willian IV	7	4	57,1%
Queen Elizabeth II	6	6	100,0%
Tudors Mary Queen	7	4	57,1%
Total	**34**	**25**	**73,5%**

Pela tabela 6.5 é possível verificar que o *recall* obtido pela ferramenta UnBWiki VSM é satisfatório, quando comparados com os resultados do estudo de caso da seção 6.1. Para o conjunto de textos o *recall* obtido é de 73,5%, tendo dois textos alcançados *recall* superior a esta marca, (textos *King Alfred the Great* e *Queen Elizabeth II*, que obtiveram *recall* de 87,5% e 100,0%, respectivamente), com destaque para o texto sobre a rainha Elizabet II cujo resultado correspondeu exatamente ao gabarito manual.

Algumas particularidades do resultado oficial do Wise 2013 contribuiram para o desempenho superior no segundo de caso, como, por exemplo, os conceitos concretos que se repetem várias vezes mas que no gabarito do Wise 2013 foram considerados apenas algumas ocorrências (conforme exemplificado na seção 6.1).

Uma outra característica que pode ter influenciado o resultado é a falta de um gabarito oficial para a análise do *Royaltree*. Como o gabarito foi gerado manualmente por um único analista, os resultados podem apresentar valores diferentes de *recall* quando comparados com a análise manual realizada por um outro analista.

Apesar das dificuldades enumeradas acima, os resultados obtidos com a metodologia ora proposta se mostraram satisfatórios, tanto em comparação com UnBWikilinks como no presente estudo de caso.

6.3 Resumo do Capítulo

Neste capítulo foram apresentados os resultados obtidos pela ferramenta UnBWiki VSM em dois estudos de caso. No primeiro deles, realizado com dados do Wise 2013, verificou-se que o *recall* obtido pela ferramenta UnBWiki VSM foi 34,2% superior ao *recall* obtido pela ferramenta UnBWikilinks, que havia alcançado resultados expressivos à época do evento. A análise efetuada levou em consideração três perspectivas: textos agrupados conforme a quantidade de menções; textos agrupados conforme o desempenho comparativo entre as duas ferramentas (*recall* pior, igual ou melhor); e textos agrupados conforme o assunto abordado. Nas três análises os resultados alcançados pela abordagem ora proposta foram superiores.

No segundo estudo de caso a ferramenta foi submetida a testes com dados reais, extraídos da página da web dedicada à história da família real britânica. O objetivo era identificar se a ferramenta ora proposta poderia ser aplicada em uma situação em que não existem links entre informações de diferentes fontes, com o objetivo de identificar conceitos concretos e nomes próprios nos textos analizados e as respectivas ligações com entidades da Wikipédia. Diante da ausência de um "gabarit" para avaliação dos resultados, a LE foi realizada previamente, de maneira manual, e o resultado obtido foi comparado com o resultado alcançado pela ferramenta UnBWiki VSM. Neste estudo de caso verificou-se que o *recall* obtido pela ferramenta, de 73,5%, foi superior ao *recall* obtido no estudo de caso anterior.

Capítulo 7

Conclusão

Esta pesquisa teve como objetivo a proposição de uma nova abordagem e o desenvolvimento de uma ferramenta para a ligação de entidades envolvendo conceitos concretos. Foram apresentadas as etapas da metodologia com foco nos dois principais problemas abordados: identificação de conceitos concretos (e nomes próprios) e seleção (desambiguação) das respectivas entidades.

Para o primeiro problema foi proposta a adoção de análise gramatical POS (*part-of-speech*) para a identificação de sentenças formadoras de conceitos concretos e nomes próprios. Conforme discutido, com base no gabarito oficial do Wise 2013 composto por 1.500 textos analisados manualmente, mais de 92% dos conceitos concretos e nomes próprios extraídos são formados por palavras pertencentes a cinco categorias distintas de POS apenas: substantivos próprios (singular e plural), substantivos comuns (singular e plural) e adjetivos. A metodologia ora proposta lança mão dessa característica para identificar, nos textos analisados, as sentenças formadoras de conceitos concretos e nomes próprios.

Com relação ao problema da seleção (desambiguação) de entidades foi proposta a adoção de uma representação baseada em Modelos de Espaço Vetorial. Nessa representação, as entidades candidatas para cada conceito concreto e/ou nome próprio são representadas como vetores em um espaço vetorial. O conceito concreto e/ou nome próprio analisado também é representado de forma vetorial no mesmo espaço (vetor-consulta), de maneira que a entidade cujo vetor estiver mais próximo do vetor-consulta apresenta a maior similaridade semântica com o conceito concreto e/ou nome próprio investigado, sendo selecionada para compor a ligação da entidade.

A metodologia foi avaliada por meio da ferramenta desenvolvida, que foi aplicada em dois estudos de caso. O primeiro estudo de caso envolveu a análise de 75 textos retirados do conjunto de textos que integrou o resultado oficial do Wise 2013, e os resultados obtidos foram comparados com o estado da arte. A metodologia ora proposta se mostrou superior, com *recall* 34,2% melhor do que o obtido na época do evento.

A segunda avaliação foi feita por meio do processamento de textos sobre a história família real britânica, disponíveis na web. Foram escolhidos aleatoriamente 5 textos que foram analisados manualmente por especialistas que identificaram os conceitos concretos e nomes próprios presentes e as respectivas entidades. Em seguida os textos foram analisados pela ferramenta ora proposta, e os conceitos concretos e nomes próprios identificados bem como as ligações de entidades efetuadas foram comparadas com os resultados obtidos de maneira manual. Neste segundo estudo de caso, o *recall* da ferramenta foi de 73,5%.

Um aspecto de melhoria detectado nos testes realizados é o tempo de processamento na análise dos textos. Nos estudos de caso foram utilizados textos de diferentes tamanhos e sobre diferentes assuntos, de maneira que a ferramenta tratou textos com muitas palavras frequentes e textos com palavras nem tão recorrentes, de forma que o tempo de processamento ocilou entre 14 e 52 minutos. Foram levantados os seguintes pontos de melhoria que podem reverter esse tempo de processamento: tratamento mais específico da base de entidades com vistas a diminuir a quantidade de palavras irrelevantes, o que provocará uma menor quantidade de cálculos no processamento das matrizes do Modelo de Espaço Vetorial; testar diferentes tecnologias de banco de dados/modelagem de dados com o objetivo de melhorar o desempenho das *queries*; e executar a ferramenta em máquinas com configuração otimizada (os testes foram realizados em uma máquina Mac OS X com processador de 2.4 GHz (Intel Core 2 Duo) e 4 GB de memória RAM.

Por outro lado, pode-se dizer que os objetivos específicos declarados no primeiro capítulo foram alcançados uma vez que: foi apresentada uma metodologia para a realização de ligação de entidades utilizando Modelos de Espaço Vetorial; foi apresentada uma metodologia para a identificação de nomes próprios e conceitos concretos com base em análise gramatical; e foi analisado o processamento de textos com o uso da ferramenta que implementou as metodologias propostas.

Outro resultado desta pesquisa é a aceitação de um artigo na edição 2015 da *International Conference on Computer and Information Technology* (IEEE CIT 2015) a ser realizado em outubro de 2015, em Liverpool, Inglaterra.

7.1 Considerações Finais e Trabalhos Futuros

A metodologia ora proposta possui aspectos que podem ser discutidos em trabalhos futuros. O primeiro deles é com relação às características da Wikipédia que foram selecionadas para a realização da seleção das entidades. Neste trabalho foram utilizados os nomes dos artigos (links) e os respectivos resumos como fonte de informação para identificação do contexto. Em um futuro trabalho podem ser utilizadas outras caracaterísticas

da Wikipédia como facilitadoras da identificação do contexto e desambiguação: páginas de redirecionamento, versões completas dos artigos, informações de mídia, etc.

Outro ponto que pode ser abordado futuramente é a aplicação da ferramenta em um contexto específico. Para se explorar as vantagens da utilização de conceitos concretos sugere-se que a ferramenta seja utilizada, por exemplo, na análise de relatórios técnicos, com o objetivo de identificar causas e soluções de problemas de maneira automática, com base nos conteúdos dos relatórios, identificando importantes fontes de conhecimento.

Referências

[1] Carolina Abreu, Flávio Costa, Laécio Santos, Lucas Monteiro, Luiz Fernando Peres de Oliveira, Patrícia Lustosa, and Li Weigang. Entity extraction within plain-text collections wise 2013 challenge - t1: Entity linking track. In *WISE (1)*, volume 8180 of *Lecture Notes in Computer Science*, pages 491–496. Springer, 2013.

[2] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 79–85, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[3] Razvan Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy*, pages 9–16, 2006.

[4] Claire Cardie. Empirical methods in information extraction. *AI magazine*, 18:65–79, 1997.

[5] Yueguo Chen, Lexi Gao, Xuan Ming, Weining Qian, and Yabo Xu. *Overview of the WISE 2013 Challenge*. Springer Berlin Heidelberg, 2013.

[6] Kenneth Ward Church. One term or two? In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 310–318, New York, NY, USA, 1995. ACM.

[7] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 249–260, Republic and Canton of Geneva, Switzerland, 2013.

[8] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, 2007. Association for Computational Linguistics.

[9] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of The American Society for Information Science*, 41(6):391–407, 1990.

[10] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). In *CIKM*, pages 1625–1628. ACM, 2010.

[11] John Rupert Firth. A synopsis of linguistic theory 1930-55. 1952-59:1–32, 1957.

[12] William Gale, Kenneth Church, and David Yarowsky. Work on statistical methods for word sense disambiguation. In *Proceedings of the AAAI Fall Symposium on Intelligent Probabilistic Approaches to Natural Language*, pages 54–60, 1992.

[13] John R. Gilbert, Cleve Moler, and Robert Schreiber. Sparse matrices in MATLAB: Design and implementation. *SIAM J. Matrix Anal. Appl.*, 13(1):333–356, 1992.

[14] Ben Hachey, Will Radford, and James R. Curran. Graph-based named entity linking with Wikipedia. In *Proceedings of t he 12th International Conference on Web Information System Engineering*, pages 213–226, Berlin, Heidelberg, 2011. Springer-Verlag.

[15] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating entity linking with Wikipedia. *Artif. Intell.*, 194:130–150, 2013.

[16] Xianpei Han and Le Sun. A generative entity-mention model for linking entities with knowledge base. In *ACL*, pages 945–954. The Association for Computer Linguistics, 2011.

[17] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[18] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226, New York, NY, USA, 2006. ACM.

[19] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[20] William P. Jones and George W. Furnas. Pictures of relevance: A geometric analysis of similarity measures. *J. Am. Soc. Inf. Sci.*, 38(6):420–442, 1987.

[21] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457–466, New York, NY, USA, 2009. ACM.

[22] George Lakoff. *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago, 1987.

[23] Landauer. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007.

[24] Thomas K Landauer and Susan T. Dutnais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[25] Dekang Lin and Patrick Pantel. DIRT - discovery of inference rules from text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328, 2001.

[26] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[27] Will Lowe. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, pages 576–581, 2001.

[28] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments and Computers*, 28 (2):203–208, 1996.

[29] Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[30] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[31] Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie Strassel. An evaluation of technologies for knowledge base population. In *LREC*. European Language Resources Association, 2010.

[32] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8, New York, NY, USA, 2011. ACM.

[33] Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, pages 233–242, New York, NY, USA, 2007. ACM.

[34] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[35] Maria Milosavljevic, Jean-Yves Delort, Ben Hachey, Bavani Arunasalam, Will Radford, and James R. Curran. Automating financial surveillance. *User Centric Media*, pages 305–311, 2010.

[36] Preslav Nakov and Marti Hearst. Solving relational similarity problems using the Web as a corpus. Columbus, OH, 2008.

[37] Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33, 2007.

[38] Simone Paolo Ponzetto and Michael Strube. Taxonomy induction based on a collaboratively built knowledge repository. *Artificial Intelligence*, 175:1737–1756, 2011.

[39] Reinhard Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322, 2003.

[40] Cornelis J. van Rijsbergen. *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA, 2004.

[41] Gerard Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[42] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA, 1987.

[43] Gerard Salton, Andrew Wong, and Chung S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[44] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[45] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[46] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. Linden: Linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 449–458, New York, NY, USA, 2012. ACM.

[47] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, New York, NY, USA, 1996. ACM.

[48] Edward E. Smith, Daniel N. Osherson, Lance J. Rips, and Margaret Keane. Combining prototypes: A selective modification model. *Cognitive Science*, 12(4):485–527, 1988.

[49] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM.

[50] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CONLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, pages 142–147, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[51] Peter D. Turney. Similarity of semantic relations. *Comput. Linguist.*, 32(3):379–416, 2006.

[52] Peter D. Turney. The latent relation mapping engine: Algorithm and experiments. *J. Artif. Intell. Res. (JAIR)*, 33:615–655, 2008.

[53] Peter D. Turney and Jeffrey Bigham. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489, 2003.

[54] Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, 2010.

[55] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[56] Mohamed Amir Yosef, Johannes Hoffart, Ilaria Bordino, Marc Spaniol, and Gerhard Weikum. AIDA: an online tool for accurate disambiguation of named entities in text and tables. *PVLDB*, 4(12):1450–1453, 2011.

[57] Mu Zhu. Recall, precision and average precision - technical report. Departament of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2004.