UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Fábio Bif Goularte

Mineração e uso de padrões linguísticos para desambiguação de palavras e análise do discurso

Florianópolis
2020

Fábio Bif Goularte

Mineração e uso de padrões linguísticos para desambiguação de palavras e análise do discurso

Tese submetida ao Programa de Pós-graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de doutor em Ciência da Computação.
Orientador: Prof. Renato Fileto, Dr.
Coorientadora: Silvia Modesto Nassar, Dra.

Florianópolis
2020

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Goularte, Fábio Bif
Mineração e uso de padrões linguísticos para desambiguação de palavras e análise do discurso / Fábio Bif Goularte ; orientador, Renato Fileto, coorientador, Silvia Modesto Nassar, 2020.
112 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2020.

Inclui referências.

1. Ciência da Computação. 2. Mineração de padrões linguísticos. 3. Extração de informação. 4. Desambiguação do sentido de palavras. 5. Análise do discurso. I. Fileto, Renato . II. Nassar, Silvia Modesto . III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

Fábio Bif Goularte

Mineração e uso de padrões linguísticos para desambiguação de palavras e análise do discurso

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Thiago Alexandre Salgueiro Pardo, Dr.
Universidade de São Paulo (videoconferência)

Prof. Horacio Saggion, Dr.
Universitat Pompeu Fabra (videoconferência)

Prof. Alckmar Luiz dos Santos, Dr.
Universidade Federal de Santa Catarina

Prof. Ricardo Azambuja Silveira, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Ciência da Computação.

Documento assinado digitalmente
Vania Bogorny
Data: 17/04/2020 12:34:47-0300
CPF: 684.586.570-15

Vania Bogorny, Dra.
Coordenadora do Programa

Documento assinado digitalmente
Renato Fileto
Data: 17/04/2020 15:56:38-0300
CPF: 106.315.998-94

Prof. Renato Fileto, Dr.
Orientador

Silvia Modesto Nassar, Dra.
Coorientadora

Florianópolis, 11 de Fevereiro de 2020.

Este trabalho é dedicado aos meus familiares e amigos,
em especial à minha mãe, Maria Bif Goularte.

AGRADECIMENTOS

Gostaria de primeiramente registrar meus agradecimentos às oportunidades e a todas as pessoas que conheci no decorrer deste período de doutorado. Embora o desenvolvimento desta tese tenha ocorrido efetivamente no último ano do doutorado, sua realização iniciou há alguns anos, e somente foi possível graças aos esforços de várias pessoas.

Agradeço à minha família, minha mãe Dona Maria, minhas irmãs Edena, Dilma e meus sobrinhos por todo amor, dedicação, apoio, oração, solidariedade e compreensão pelas ausências justificadas. Aos meus amigos que em muitos momentos foram à minha família.

Agradeço aos meus orientadores. Ao professor Renato Fileto pela orientação acadêmica, profissional e pessoal, pela atenção, pela amizade, e por sempre procurar extrair o melhor de mim nos estudos, acreditando em meu potencial e ajudando na minha formação como pesquisador. À professora Silvia Modesto Nassar pela orientação, pela amizade, incentivo, humildade, pela sua humanidade e caráter, e por sempre procurar entender as características individuais de cada um de seus orientados ("filhos acadêmicos"), fazendo disso uma potencialidade, e não um problema. Acredito que consegui convergir um pouco do conhecimento destes dois grandes professores nesta tese.

Agradeço ao professor Horacio Saggion, da Universitat Pompeu Fabra (UPF), e a todos os integrantes do Tractament Automàtic del Llenguatge Natural (TALN group), que me receberam de maneira amigável e tornaram minha estadia em Barcelona muito proveitosa. Agradeço à Danielly Sorato, pela parceria em pesquisas e trabalhos.

Agradeço a todos que contribuíram, direta ou indiretamente, para o desenvolvimento deste trabalho. À Universidade Federal de Santa Catarina, aos professores e funcionários do INE-UFSC e PPGCC-UFSC, aos membros das bancas do exame de qualificação e da defesa. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que por meio de bolsa de estudo (001 e 88881.135663/2016-01), apoiou financeiramente a realização desta pesquisa.

Por fim, agradeço aos desconhecidos, a Deus, cujas obras e os feitos me libertaram e me motivaram a continuar.

RESUMO

A extração de informação contida em textos na web tem o potencial de alavancar uma série de aplicações, mas muitas delas requerem a captura automática da semântica exata de elementos textuais relevantes. O Twitter, por exemplo, gera diariamente centenas de milhões de pequenos textos (*tweets*), muitos dos quais com rica informação sobre usuários, fatos, produtos, serviços, desejos, opiniões, etc. A anotação semântica de palavras relevantes em *tweets* é um grande desafio, pois eles impõem dificuldades adicionais (e.g., pouca informação de contexto, agramaticalidade) para métodos automáticos realizarem uma desambiguação de qualidade, o que leva a resultados com baixa precisão e cobertura. Inclusive, porque a língua é um sistema simbólico polissêmico, que não tem uma semântica pronta, o que se manifesta acentuadamente em linguagem coloquial e particularmente em mídias sociais. As soluções atuais de anotação geralmente não conseguem encontrar o sentido correto de palavras em construções envolvendo a semântica implícita que, às vezes, é colocada intencionalmente, por exemplo, para fazer humor, ironia, jogo de palavras ou trocadilhos. Este trabalho propõe o desenvolvimento de uma abordagem para minerar padrões léxico-semânticos, com a finalidade de captar a semântica em texto para utilizar em tarefas que processam a linguagem. Estes padrões foram denominados de padrões MSC^+, pois são definidos por sequências de Componentes Morfo-semânticos (MSC). Um algoritmo não-supervisionado foi desenvolvido para minerar tais padrões, que suportam a identificação de um novo tipo de característica semântica em documentos, assim como métodos para desambiguar o sentido de palavras. Os resultados de experimentos com a tarefa de Word Sense Disambiguation (WSD), em texto de mídia social, mostraram que instâncias de alguns padrões MSC^+ aparecem em vários tweets, mas às vezes usando palavras diferentes para transmitir o sentido. Os testes realizados nos resultados do experimento em WSD demonstraram que a exploração dos padrões MSC^+ permite mecanismos eficazes na desambiguação do sentido de palavras, levando a melhorias no estado da arte, segundo medidas de precisão, cobertura e medida-F. Os padrões MSC^+ também foram explorados em experimentos com Análise do Discurso (AD) do conteúdo de diferentes obras do escritor Machado de Assis. Os experimentos revelaram a incidência de padrões morfo-semânticos que evidenciam características de obras literárias e que podem auxiliar na classificação de discurso das obras analisadas, tais como a preponderância de verbos específicos nos contos, de substantivos femininos nos romances e adjetivos nos poemas.

Palavras-chave: Extração de informação. Sumarização de tweets. Padrões linguísticos. Desambiguação do sentido de palavras. Análise do discurso.

ABSTRACT

Information extraction from social media texts has the potential to boost a number of applications, but many of them require the automatic capture of accurate semantics of relevant textual elements. Twitter, for example, generates hundreds of millions of short texts (tweets) daily, many of which containing rich information about users, facts, products, services, desires, opinions, etc. The semantic annotation of relevant words in tweets is a challenge because social media impose additional difficulties (e.g., little context information, poor grammatical rules conformity) for automatic methods to carry out quality disambiguation. It leads to results with low accuracy and coverage. In addition, a language is a polysemic symbolic system without ready semantics for some constructs. Sometimes words have implicit semantics (e.g., to make humor, irony, wordplay). It is common in colloquial language, and particularly in social media. In this work, we propose the development of an approach to mine lexical-semantic patterns and capture the semantics of texts for use in language processing tasks. We learn these patterns, that we call MSC^+ patterns, from text data defined by Morpho-semantic Components (MSC). An unsupervised algorithm was developed to mine such patterns, which support the identification of a new kind of semantic feature in documents, as well as methods for disambiguating the meaning of words. Experimental results on Word Sense Disambiguation (WSD) task, from tweets, show that instances of some MSC^+ patterns arise in many tweets, but sometimes using different words to convey the sense of the respective MSC in some tweets where pattern instances appear. The exploitation of MSC^+ patterns when they induce semantics on target words enables effective word sense disambiguation mechanisms leading to improvements in the state of the art (e.g., metrics such as accuracy, coverage, and F-measure). We also explored the MSC^+ patterns on the Discourse Analysis (DA) with literary content. Experimental results on selected works of a Brazilian writer submitted to our algorithm reveal the incidence of distinct morpho-semantic patterns in different types of works, such as the preponderance of specific verbs in tales, feminine nouns in romances, and adjectives in poems.

Keywords: Information extraction. Tweet summarization. Linguistic patterns. Word sense disambiguation. Discourse Analysis.

LISTA DE FIGURAS

Figura 1 – Exemplo de texto de mídia social (Twitter) 14
Figura 2 – O problema de WSD deste trabalho. 19
Figura 3 – Representação do contexto por meio de vetores. 24
Figura 4 – Etapas de pré-processamento e processamento do texto. 26
Figura 5 – Típico processo de enriquecimento semântico. 29
Figura 6 – Típico problema de desambiguação. 32
Figura 7 – Exemplo de extração de informação de um artigo de notícias. 41
Figura 8 – Semântica implícita em texto de mídias sociais. 43
Figura 9 – Evolução das publicações em desambiguação de palavras. 49
Figura 10 – Arquitetura da abordagem proposta. 55
Figura 11 – Abordagem proposta para WSD. 64
Figura 12 – Cálculo do escore com os métodos MAA-S e MAA-PS usando instâncias de correspondência perfeita a cada padrão candidato. . . . 71
Figura 13 – Cálculo dos escores com os métodos MAA-S e MAA-PS usando instâncias de correspondência perfeita a cada padrão candidato. . . 72
Figura 14 – Tempo de execução das amostras. 73
Figura 15 – Comparativo entre Cobertura (a) e Precisão (b). 75
Figura 16 – Abordagem proposta para AAD. 79
Figura 17 – Prevalência das anotações de classe morfossintática nos documentos 84

LISTA DE TABELAS

Tabela 2 – Significados existentes para *"polar"* na BabelNet 14
Tabela 3 – Exemplo de padrão morfo-semântico 20
Tabela 4 – Padrões de palavras frequentes 38
Tabela 5 – Exemplo de instâncias de padrões MSC^+ de comprimento 3 (MSC_1, MSC_2, MSC_3) identificados nos tweets #1 a #5 39
Tabela 6 – Estudos recuperados nas etapas de busca e seleção 49
Tabela 7 – Uma amostra de *tweets* do conjunto de dados. 65
Tabela 8 – Um exemplo de pré-processamento de *tweet*. 66
Tabela 9 – Exemplos de componentes textuais anotados. 67
Tabela 10 – Exemplo de MSC extraído de um *tweet* 68
Tabela 11 – Padrões MSC^+ e instâncias encontradas no dataset ($sup = 2\%$). . . 69
Tabela 12 – Resultados da Indução/Desambiguação do sentido 74
Tabela 13 – Resultados com padrões MSC^+ e $sup = 2\%$ (MSC 2%) 74
Tabela 14 – Resultado com padrões MSC^+ e $sup = 5\%$ (MSC 5%) 74
Tabela 15 – Teste de hipóteses sobre a significância dos ganhos com o método MAA-PS (experimentos do tipo Binomial). 75
Tabela 16 – Uma amostra de textos do conjunto de dados. 80
Tabela 17 – Um exemplo de pré-processamento de texto. 80
Tabela 18 – Exemplo de anotações de componentes em texto literário 81
Tabela 19 – Quantidades de padrões minerados e instâncias nos documentos . 83
Tabela 20 – Os três padrões MSC$^+$ mais frequentes no discurso 85

LISTA DE ABREVIATURAS E SIGLAS

AAD	Análise Automática do Discurso
AD	Análise do Discurso
AM	Aprendizado de Máquina
BCL	Base de Conhecimento Lexical
BoW	Bag-of-Words
DLS	Desambiguação Lexical do Sentido
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MAA-PS	Máxima aderência média a classe morfossintática e sentido
MAA-S	Máxima adesão média ao sentido
MSC	Morpho-Semantic Components
NED	Named Entity Disambiguation
NER	Named Entity Recognition
PLN	Processamento de Linguagem Natural
PoS	Part-of-Speech
RS	Revisão Sistemática
SED	Semantic Enrichment of Data
SSP	Short Semantic Pattern
WSD	Word Sense Disambiguation
WSI	Word Sense induction

SUMÁRIO

1 INTRODUÇÃO . . . 13
1.1 MOTIVAÇÃO . . . 15
1.2 HIPÓTESES DE PESQUISA . . . 17
1.3 OBJETIVOS . . . 18
1.4 COMPLEXIDADE E INEDITISMO . . . 19
1.5 ORGANIZAÇÃO DESTA TESE . . . 21
2 FUNDAMENTOS E DEFINIÇÃO DO PROBLEMA . . . 23
2.1 O CONTEXTO TEXTUAL . . . 23
2.1.1 Características do contexto textual . . . 25
2.2 ENRIQUECIMENTO DO CONTEXTO . . . 28
2.2.1 Enriquecimento semântico . . . 28
2.3 DESAMBIGUAÇÃO LEXICAL DE SENTIDO . . . 30
2.4 SISTEMAS WSD . . . 32
2.4.1 Fontes de conhecimento . . . 33
2.4.2 Abordagens clássicas de WSD . . . 34
2.4.2.1 Baseada em conhecimento . . . 34
2.4.2.2 Supervisionada . . . 35
2.4.2.3 Não supervisionada . . . 36
2.5 MINERAÇÃO DE TEXTO . . . 37
2.5.1 Mineração de padrões frequentes . . . 37
2.5.2 Investigação de padrões MSC^+ . . . 38
2.6 CONSIDERAÇÕES . . . 40
2.6.1 Origem do problema . . . 40
2.6.2 Mídias sociais . . . 41
2.6.3 A semântica em tweets . . . 42
2.6.4 Desafios . . . 44
2.6.5 Aplicação . . . 45
3 REVISÃO SISTEMÁTICA DA LITERATURA . . . 46
3.1 PLANO DE PESQUISA REALIZADO . . . 46
3.2 RESULTADO DA REVISÃO SISTEMÁTICA . . . 48
3.3 MINERAÇÃO DE PADRÕES LINGUÍSTICOS E APLICAÇÕES . . . 52
4 ABORDAGEM MSC^+ . . . 55
4.1 FORMALIZAÇÃO . . . 56
4.1.1 Padrão MSC^+ . . . 58
4.1.2 Mineração de padrões MSC^+ . . . 59
4.1.3 WSD por meio de padrões MSC^+ . . . 60
5 USO DE PADRÕES MSC^+ EM WSD . . . 63

5.1 DESCRIÇÃO DO EXPERIMENTO 63
5.1.1 Base de dados utilizada 65
5.1.2 Pré-processamento e sumarização de *tweets* 65
5.1.3 Anotação morfossintática e de sentidos candidatos 66
5.1.4 Consolidação de anotações 67
5.1.5 Extração de MSCs 68
5.1.6 Mineração de padrões MSC^+ 69
5.1.7 WSI/WSD com padrões MSC^+ 69
5.2 RESULTADOS 72
5.2.1 Tempo de execução 72
5.2.2 Indução/Desambiguação do sentido 73
5.3 CONSIDERAÇÕES 76
6 USO DE PADRÕES MSC^+ EM ANÁLISE MORFO-SEMÂNTICA DE DISCURSOS 78
6.1 DESCRIÇÃO DO EXPERIMENTO 78
6.1.1 Base de dados utilizada 79
6.1.2 Pré-processamento 80
6.1.3 Anotação morfossintática e semântica 80
6.1.4 Consolidação de anotações 82
6.1.5 Extração de MSCs 82
6.1.6 Mineração de padrões MSC^+ e seleção de instâncias 82
6.2 RESULTADOS 83
6.3 CONSIDERAÇÕES 86
7 CONCLUSÃO 87
7.1 PRINCIPAIS CONTRIBUIÇÕES 87
7.2 CONTRIBUIÇÕES CIENTÍFICAS 88
7.3 PUBLICAÇÕES 91
7.4 LIMITAÇÕES E TRABALHOS FUTUROS 93
REFERÊNCIAS 95

1 INTRODUÇÃO

O significado ou a semântica de uma sentença é fundamental na comunicação e aquisição de conhecimento (KUMAR, 2015). Como a linguagem natural é ambígua, uma palavra pode assumir diferentes significados dependendo do contexto em que é utilizada (NAVIGLI, 2009). O significado de palavras, simples ou compostas, pode ser representado por uma definição, uma lista de sinônimos (*synset*) ou um exemplo de senso de uso comum. O processo de atribuir o sentido correto a uma palavra é chamado de desambiguação de sentido, e o estudo desse problema no campo da computação é conhecido pelo nome de Desambiguação Lexical do Sentido (DLS) ou *Word Sense Disambiguation (WSD)* (JURAFSKY; MARTIN, 2009; AGIRRE; EDMONDS, 2007).

Existem duas abordagens principais para WSD, a supervisionada e a não supervisionada. A abordagem supervisionada é baseada em Aprendizado de Máquina (AM) (e.g., classificação, reconhecimento de padrões) e necessita de uma grande coleção de dados de treinamento para construir o modelo de classificação (NAVIGLI, 2009). Geralmente, a construção dos dados de treinamento é uma tarefa onerosa porque envolve a anotação manual de palavras por especialistas. Já a abordagem não supervisionada usa dados de um repositório externo (e.g., grandes volumes de textos disponíveis, padrões linguísticos encontrados em tais textos) para determinar o exato sentido das palavras. O repositório de sentidos para as palavras (AGIRRE; EDMONDS, 2007) pode ser um dicionário de *synsets* com definições e exemplos de uso, uma ontologia de domínio ou uma rede semântica, produzida a partir de coleções de dados ligados. Esses repositórios exigem atualização periódica. Se uma nova palavra é adicionada à rede semântica, novos sentidos também são introduzidos e isso pode afetar toda a rede.

Atualmente, existem vários repositórios com informações sobre o sentido de palavras. Os mais utilizados em trabalhos sobre Semântica[1] são WordNet (MILLER, G. A. *et al.*, 1990; FELLBAUM, 1998) e BabelNet (NAVIGLI; PONZETTO, 2010). Ambas apresentam uma ampla rede semântica de cobertura do sentido de palavras. Para a língua Inglesa, a versão original da WordNet[2] inclui mais de 206 mil formas de sentidos e mais de 155 mil palavras, enquanto que a BabelNet[3] tem mais de 22 milhões formas de sentidos e mais de 3 milhões de conceitos e 5 milhões de entidades nomeadas. Entretanto, apesar das amplas coberturas dessas bases, muitas palavras e sentidos estão omitidos (e.g., domínios específicos como o de leis ou medicina, uso de gírias criativas para entidades que surgem em um momento recente, palavras que assumem

[1] International Workshop on Semantic Evaluation - SemEval, `https://aclweb.org/aclwiki/SemEval_Portal`

[2] `https://wordnet.princeton.edu/documentation/wnstats7wn`

[3] `http://babelnet.org/stats`

um novo valor semântico em consequência do contexto).

Para ilustrar o problema tratado nesta pesquisa, a Figura 1 apresenta um texto de mídia social (*tweet*) e a Tabela 2 mostra os significados encontrados para a palavra alvo "*polar*" na rede BabelNet (5 sentidos para substantivos e 2 para adjetivos).

Figura 1 – Exemplo de texto de mídia social (Twitter)

Tabela 2 – Significados existentes para *"polar"* na BabelNet

#	Synset	Classe	Significado
1	Polar (Crash Bandicoot)	substantivo	Esta é uma lista de personagens da série de jogos.
2	Polar (Satélite)	substantivo	O Polar Satellite é um satélite lançado pela NASA.
3 •	Polar (estrela variável, cataclísmica)	substantivo	As Polares são uma classe de variáveis cataclísmicas magnéticas.
4 ⋆	Polar (cerveja, Cervejaria Polar, Polar Export)	substantivo	Polar Export é uma marca de cerveja pilsen brasileira.
5	Polar (cola), Cola Polar	substantivo	Cola da empresa brasileira Compactor.
6	polar	adjetivo	Próximo ao polo geográfico (tradução).
7	gelado, polar, ártico	adjetivo	Extremamente frio.

⋆ Significado correto; • Significado desambiguado pela ferramenta Babelfy.

O texto foi submetido à ferramenta Babelfy[4] que fez a desambiguação e anotou "*polar*" com o sentido presente na linha 3 da Tabela 2. Todavia, para o contexto do texto o sentido correto é o que está na linha 4. Também foi observado equívoco na anotação gerada pelo Babelfy para o verbo "*tomar*". Além disso a palavra "*trincando*" não existe na BabelNet. A palavra "*trincando*" deriva do verbo trincar, mas no contexto do texto ela assume o valor de um adjetivo com o sentido de "*muito gelado*". A desambiguação correta de palavras em textos curtos, como "*polar*" e "*tomar*" e, a captura do significado

[4] `http://babelfy.org/index`, acessado em maio de 2018.

correto de "*trincando*", representa um grande desafio para ferramentas de anotação automática. A anotação dessas palavras envolve semântica implícita, a qual requer informação de contexto (textual e cultural) para ser apreendida.

Várias técnicas têm sido propostas para ampliar a estrutura e enriquecer a WordNet e a BabelNet com novas palavras e sentidos (SNOW; JURAFSKY; NG, 2006; TORAL; MONACHINI, 2008; PONZETTO; NAVIGLI, 2009; NAVIGLI; PONZETTO, 2010; JURGENS; PILEHVAR, 2015; GREGORY; RAVELLI; PANUNZI, 2016). O enriquecimento de tais fontes melhora a cobertura dos sentidos, mas também pode tornar a tarefa de sistemas de WSD mais difícil. Novas definições de sentidos podem não corresponder ao estilo de escrita, ao cumprimento de normas da língua ou ao idioma utilizado para os sentidos já armazenados na fonte. Esses são alguns dos fatores que influenciam na baixa precisão de sistemas de WSD. Além disso, definir o nível de detalhamento de um sentido em relação ao uso (e.g., mais fino ou mais genérico) ou a quantidade de contexto necessária para representá-lo também pode ser um problema (NAVIGLI, 2009).

1.1 MOTIVAÇÃO

Os seres humanos são capazes de entender o significado de um determinado texto porque usam vários recursos (e.g., conhecimento prévio, informação implícita, inferência) para buscar a melhor definição de cada palavra. Por outro lado, o desempenho de técnicas de WSD ainda está longe daquele dos humanos (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016), principalmente se o texto apresentar uma semântica desafiadora, como acontece em certos documentos de bibliotecas digitais (e.g., textos literários), ambientes virtuais de aprendizagem (e.g., textos de estudantes) e mídias sociais (ANTONIOU; VAN HARMELEN, 2004). Textos curtos também dificultam a anotação semântica devido ao contexto limitado. Por exemplo, métodos de reconhecimento de entidade nomeadas, tipicamente apresentam $85 - 90\%$ de precisão em textos mais longos, mas $30 - 50\%$ em *tweets* (LIU, X. *et al.*, 2012; BONTCHEVA; DERCZYNSKI, 2016). Também é possível verificar que a precisão de métodos para WSD varia de $20 - 30\%$ conforme o domínio e o idioma (MORO; CECCONI; NAVIGLI, 2014).

Textos de mídias sociais impõem dificuldades adicionais para métodos automáticos realizarem uma anotação semântica de qualidade, contribuindo para resultados com baixa precisão e cobertura. Dentre os principais desafios impostos por tais textos às ferramentas de anotação automática, pode-se citar o uso de marcações específicas (e.g., *@usuário*, *URIs*, *hashtags*), a pouca informação de contexto, a presença de ruído, a semântica mal definida, o emprego de linguagem agramatical (LAENDER *et al.*, 2002; SORATO; GOULARTE *et al.*, 2016).

A criação automática de anotações semânticas de qualidade em textos de mídias sociais tem o potencial de alavancar uma série de aplicações. A tarefa de

WSD pode, por exemplo, associar a componentes linguísticos relevantes semântica bem definida referente a eventos ocorridos ao longo do espaço e do tempo (WANG; STEWART, 2015; ANANTHARAM *et al.*, 2015; XIA *et al.*, 2015; HUA; ZHENG; ZHOU, 2015), locais frequentemente visitados (SUVITHA; JANARTHANAN, 2012; FILETO *et al.*, 2015), pessoas, instituições, produtos, serviços, etc (MENDES; JAKOB *et al.*, 2011; MORO; RAGANATO; NAVIGLI, 2014). Isso permite alimentar com semântica precisa métodos de análise de informação (SACENTI *et al.*, 2015), análise de sentimento (DAS; ACHARJYA; PATRA, 2014), mineração de dados (LINOFF; BERRY, 2011), recomendação (PAZZANI; BILLSUS, 2007), geração de respostas (i.e. Question answering Systems (QaS)) (AL-HARBI; JUSOH; NORWAWI, 2017), simplificadores de texto (SIDDHARTHAN; MANDYA, 2014), sumarizarizadores (RUDRA *et al.*, 2016), tradução automática (JONES *et al.*, 2012) e bases de conhecimento (MILLER, G. A. *et al.*, 1990; SCARTON, 2013; CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016)

Esta tese pretende contribuir com a área de estudo em WSD propondo uma abordagem não supervisionada para desambiguar o sentido de palavras em textos de mídias sociais (postagens no Twitter). A abordagem proposta combina a análise de níveis de processamento da língua (morfológico, sintático e semântico), sumarização de texto e bases de conhecimento para desambiguar possíveis significados em textos. Por esse motivo, a abordagem proposta também acrescenta contribuições a áreas de estudo que não são alvo de investigação, como análise morfo-semântica de discurso.

O significado de palavras pode ser indiretamente inferido a partir do conteúdo dos textos, mediante exploração de padrões linguísticos. **Padrões linguísticos** são modelos e regras linguísticas que permitem a seus usuários se comunicarem adequadamente usando uma linguagem comum (SILVA, 2017). A língua e suas variações são estruturas regidas por gramáticas finitas, que usam recursos léxicos infinitos (i.e. teoria da gramática gerativa) (CHOMSKY, 1959). Do ponto de vista da linguística, uma gramática não é apenas uma coleção de regras, mas um conjunto de modelos que orientam os falantes a produzir frases compreensíveis e previsíveis, e que consiste em padrões que fazem sentido com o uso de símbolos, sons e palavras arbitrárias que compõem a língua (DECAPUA, 2016).

Um padrão linguístico define dois aspectos em relação ao uso de uma linguagem controlada: (i) **vocabulário**, com um conjunto limitado de termos incluindo a semântica dos termos; e (ii) **sintaxe**, com um conjunto controlado de atributos de cada elemento sintático com o respectivo nome, tipo e multiplicidade (e.g., obrigatório ou opcional). No entanto, padrões linguísticos podem ser representados de diferentes formas.

Nesta tese, um padrão linguístico é representado por estilos linguísticos, definidos por alguma linguagem natural controlada, mais precisamente por Padrões Morfo-Semânticos (MSC^+). Um padrão MSC^+ refere-se a um conjunto ordenado de fragmentos de texto, cada qual com dois aspectos relevantes: sintático (marcação de *PoS*)

e semântico (sentido candidato). Cada padrão MSC^+ é representado por um conjunto de pares $\langle PoS, Sentido \rangle$, que correspondem à classificação de *Part-of-Speech* (*PoS*) seguida pelo sentido atribuído ao fragmento de texto (obtido a partir de um conjunto de classes de sentido, ontologia, base de dados com léxicos controlados, dados ligados). Por exemplo, o texto de *tweet* apresentado na Figura 1 segue um padrão MSC^+ que pode ser expresso pelo seguinte conjunto de componentes morfo-semânticos $\langle Verbo, Ingerir \rangle$, $\langle Sustantivo, Cerveja \rangle$, $\langle Adjetivo, Qualquer \rangle$. Outros textos de *tweets* com instâncias deste padrão são apresentados na Tabela 3 (Secão 1.4), onde os três componentes morfo-semânticos são realçados nas colunas w_1, w_2, w_3, respectivamente.

O estado da arte em Processamento de Linguagem Natural (PLN) monstra que com a evolução de máquinas e softwares, modelos matemáticos ou genéricos passaram a ser utilizados nos diferentes níveis de processamento da linguagem para transformar, inferir e manipular textos. Para desenvolver estes modelos é necessário combinar tarefas de aprendizado dos dados a outras teorias de representação de conhecimento, o que exige esforço de construção, conhecimento especializado e dados apropriados para treinamento e validação do modelo. Além disso, há casos em que, devido à falta de transparência no funcionamento do modelo, não se consegue ter clareza sobre os resultados. Neste sentido, com o entendimento de que esta tese apresenta uma nova investigação empírica sobre padrões linguísticos (padrão MSC^+), a abordagem proposta possui características de processamento simbólico, pois se baseia em regras linguísticas bem estruturadas e sem ambiguidades. Isso contribui para a explicabilidade do modelo proposto que, por sua vez, não requer grandes conjuntos de dados anotados, como em abordagens de aprendizado profundo.

1.2 HIPÓTESES DE PESQUISA

Muito já se sabe sobre a tarefa de WSD, mas poucas pesquisas são destinadas especificamente a textos de mídias sociais, principalmente em língua Portuguesa (NÓBREGA, 2013; SCARTON, 2013). Além disso, há categorias de palavras, como os verbos, e estilos de texto que necessitam de uma descrição linguística mais detalhada (GUREVYCH; ECKLE-KOHLER; MATUSCHEK, 2016). O nível de complexidade dos textos, nesses casos, dificulta o processamento e consequentemente suas minúcias de significação não são captadas. Em virtude disso, ferramentas de anotação semântica, tais como Babelfy (MORO; RAGANATO; NAVIGLI, 2014) e DBpedia Spotlight (MENDES; JAKOB *et al.*, 2011), ainda apresentam uma baixa eficiência. Esse problema é devido principalmente à ausência de traços que vão além dos níveis léxico e morfossintático, para considerar os semântico-discursivos. Assim, esta tese parte das seguintes hipóteses de pesquisa:

1. **Técnicas de sumarização de texto podem detectar sequências de postagens de mídias sociais com semântica mais importante**. Textos, especialmente textos coloquiais como postagens em mídias sociais, podem ter muitos componentes semanticamente relevantes, mas que precisam ser identificados.

2. **É possível explicitar automaticamente a semântica implícita**. A significação de palavras não costuma ser única nem estática. As palavras podem ter seu significado modificado dependendo do contexto, deixando de representar apenas a ideia original (básica e objetiva), remetendo-nos a novos sentidos por meio de associações, às vezes, indiretas (semântica implícita). As informações de contexto do texto, incluindo as do autor e as de populações com os traços culturais análogos, podem ser utilizadas para desambiguar ou mesmo induzir o sentido de tais palavras, em várias situações em que isso se faz necessário.

3. **É possível melhorar a precisão e cobertura de métodos de anotação semântica por meio de padrões linguísticos**. Com a utilização de estruturas semânticas de bases de conhecimento, juntamente com informação de contexto, pode-se extrair a semântica implícita de textos de mídias sociais, e assim captar o significado de palavras que métodos de anotação semântica atuais não fazem.

1.3 OBJETIVOS

Tendo em vista as hipóteses definidas, esta tese tem o objetivo de **propor uma abordagem para minerar padrões linguísticos em texto, explorando as relações semânticas de palavras relevantes em *tweets* e textos em língua Portuguesa, para determinar o significado de palavras e analisar o conteúdo do discurso**. A principal novidade da técnica proposta é a mineração de padrões MSC^+, que se baseia em uma nova forma de representação semântica de texto definida por Componentes Morfo-Semânticos (*Morpho-Semantic Components - MSC*), além de sua aplicação em WSD e Análise do Discurso. Como consequência do objetivo apresentado foram definidos objetivos específicos que evidenciam as principais contribuições desta tese:

1. Descrever e detalhar o estado da arte sobre técnicas e ferramentas de anotação semântica automática, principalmente daquelas para reconhecimento e classificação de palavras relevantes e desambiguação do sentido de palavras.

2. Selecionar ferramentas, bases de dados e bases de conhecimento para realizar a identificação de palavras relevantes, classificação morfossintática e anotação semântica, visando auxiliar no processo de captura de sentidos corretos de palavras.

3. Desenvolver a definição de Componente Morfo-semântico (MSC), Padrão Morfo-semântico (MSC^+), o algoritmo para a mineração de tais padrões em texto e métodos para desambiguar o sentido de palavras.

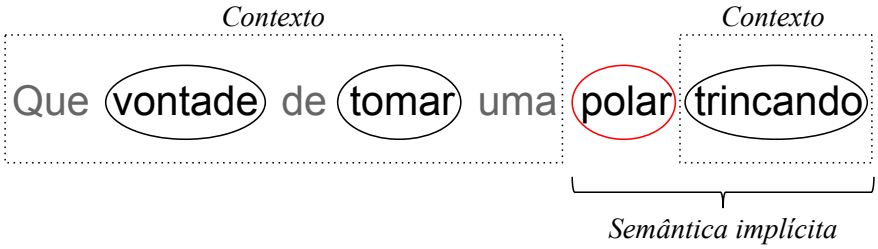
4. Implementar uma abordagem, preferencialmente multilingue e aplicável a diferentes conjuntos de dados, para captar a semântica implícita de palavras relevantes em sequências de textos curtos, com base em técnicas de sumarização de texto e análise do contexto.

5. Testar a abordagem proposta em experimentos com dados de microblogs, e textos com ênfase em Português do Brasil, para comparar a quantidade e a qualidade dos sentidos atribuídos com aqueles de técnicas do estado da arte.

1.4 COMPLEXIDADE E INEDITISMO

O exemplo a seguir ilustra de maneira mais concreta a problemática tratada e o que se pretende fazer. A sentença da Figura 2 *"Que vontade de tomar uma polar trincando"* (extraída do Twitter) apresenta palavras que são relevantes (circuladas) e outras que podem ser chamadas de *stopwords*. Para desambiguar a palavra alvo "*polar*", podem-se usar as palavras relevantes próximas do alvo (contexto textual) para ajudar a entender o sentido correto.

Figura 2 – O problema de WSD deste trabalho.



Os significados das palavras relevantes podem corresponder aos de *synsets* da WordNed, VerbNet.Br ou OpenWN-PT (versões em português dos originais em inglês da VerbNet e da WordNet, respectivamente). Eles também podem ser equivalentes a recursos de coleções de dados ligados como a DBpedia ou nodos da BabelNet, que integra recursos léxicos e dados ligados. A partir dos melhores significados aferidos em bases de conhecimento, pretende-se captar a semântica de palavras alvo existentes na BabelNet.

Podem-se explorar dois tipos de representações de contextos para abordar a semântica implícita em textos: baseados em redes semânticas e em padrões linguísticos.

Os contextos baseados em redes semânticas (D'AGOSTINI; FILETO, 2009) capturam a semântica usualmente utilizada por um autor ou grupo de autores por meio de pesos atribuídos a nodos, arestas ou porções de uma rede semântica. Os métodos baseados em contextos semânticos visam a maximização da coerência semântica e estão sendo aperfeiçoados e aplicados à desambiguação de anotações em outros trabalhos do nosso grupo de pesquisa (GBD/UFSC).

O contexto baseado em redes semânticas também pode ajudar a expressar o contexto em padrões linguísticos. Por exemplo, para desambiguar a palavra "*polar*" para o sentido relacionado à marca de cerveja, pode-se utilizar casos em que um autor e/ou seus interlocutores mencionem com frequência coisas relacionadas a bebidas ou tenham estado recente ou frequentemente em local(ais) como um bar.

O contexto baseado em padrões linguísticos, por outro lado, é explorado no âmbito deste trabalho. Basicamente, esse contexto captura formas típicas de expressão, como sequências de palavras que aparecem em textos, cada qual com uma classe morfossintática e um significado. Usando o contexto baseado em padrões linguísticos pretende-se desambiguar, ou mesmo induzir, o significado de certas palavras em uma instância do padrão. A indução de sentido pode ser com base no significado usualmente atribuído à palavras correspondentes em outras instâncias do mesmo padrão.

A Tabela 3 mostra quatro instâncias de textos curtos (#) onde se observa um padrão morfo-semântico, definido por uma sequência de três palavras (w_1, w_2, w_3), sendo: w_1 um verbo (com sentido próximo a querer, pedir ou ingerir), w_2 um substantivo (com sentido de bebida ou comida) e w_2 um adjetivo (com sentido de gelada).

Tabela 3 – Exemplo de padrão morfo-semântico

#		w_1		w_2		w_3
1	Quero	**tomar**	uma	**cerveja**	bem	**gelada**.
2	Vontade de	**tomar**	aquela	**cerveja**	super	**gelada**.
3	Vendo jogo	**bebendo**	uma	**brahma**		**gelada**.
4	Vontade de	**tomar**	uma	**polar**		**trincando**.

A semântica armazenada em bases de conhecimento, às vezes, não é suficiente para expressar todos os significados possíveis do texto de mídias sociais. Por exemplo, a palavra "*trincando*" na Figura 2, que está sendo utilizada com o sentido de um adjetivo, mas é originalmente um verbo. Entretanto, é possível apoiar-se na semântica desambiguada da palavra correspondente em outras sentenças que seguem o mesmo padrão morfossintático para induzir o significado correto. Todavia, preferiu-se deixar fora do escopo desta tese investigações adicionais sobre indução de significado, tais como casos que não podem ser solucionados pelo método aqui proposto, consistência e impactos em bases de conhecimento, inclusive porque esses temas estão sendo estudados por outros pesquisadores.

1.5 ORGANIZAÇÃO DESTA TESE

Este documento está organizado da seguinte forma:

Capítulo 2 - Fundamentos e definição do problema. Neste capítulo é apresentada a fundamentação teórica sobre os temas que envolvem o problema deste estudo: a semântica, principalmente a existente em texto de microblogs, o enriquecimento semântico de texto, a definição da tarefa de WSD, a mineração de texto e as explorações iniciais sobre Padrões MSC^+. A intenção é fornecer ao leitor um entendimento aprofundado sobre o tema estudado nesta tese e tornar a conteúdo desenvolvido acessível a um público mais amplo.

Capítulo 3 - Revisão sistemática da literatura. Neste capítulo é apresentada a revisão da literatura sobre o tema desta tese. A revisão da literatura foi realizada de forma sistemática, seguindo um protocolo para responder algumas questões de pesquisa levantadas.

Capítulo 4 - Padrões MSC^+. Neste capítulo é apresentada a abordagem proposta baseada em padrões linguísticos denominados MSC^+, os quais são expressos por sequências de Componentes Morfo-Semânticos (*Morpho-Semantic Components (MSC)*).

Capítulo 5 - Uso de padrões MSC^+ em WSD. Neste capítulo é apresentado um plano de experimento para validação da abordagem proposta na tarefa WSD. É apresentada uma sequência de etapas que compõem o processo da tarefa WSD baseada na abordagem MSC^+ para texto de mídia social. Também são apresentados os resultados de cada etapa, com exemplos, assim como a seleção de ferramentas, bases de dados e tarefas de processamento da linguagem. Por fim, este capítulo apresenta a discussão dos resultados do experimento e considerações.

Capítulo 6 - Uso de padrões MSC^+ em análise morfo-semântica de discursos. Neste capítulo é descrito um experimento para teste da abordagem proposta na tarefa de Análise do Discurso/Conteúdo. Semelhante ao experimento com WSD, neste experimento é apresentada uma sequência de etapas que compõem um processo de descoberta de características sobre o discurso de algumas obras literárias. Na sequência, este capítulo apresenta a análise e discussão dos resultados sobre os traços identificados no discurso das obras.

Capítulo 7 - Conclusão. Neste capítulo são apresentadas as considerações finais deste trabalho. Alguns itens descritos na Introdução (hipóteses e objetivos), assim como as questões de pesquisa presentes no Capítulo 3, são retomados e

discutidos, apontando as principais contribuições científicas. Na sequência, são apresentadas as publicações relacionadas a esta tese e, por fim, as limitações e trabalhos futuros.

2 FUNDAMENTOS E DEFINIÇÃO DO PROBLEMA

Neste capítulo são abordados alguns dos fundamentos sobre os quais esta tese foi desenvolvida. Inicialmente são apresentados os principais conceitos utilizados no trabalho. A Seção 2.1 apresenta uma discussão sobre o contexto na caracterização da semântica das palavras. Logo após, a Seção 2.2 descreve os principais tipos de enriquecimento de léxicos, tendo o foco no enriquecimento semântico. A Seção 2.3 define formalmente o problema de WSD e delimita o escopo desta pesquisa. A Seção 2.4 apresenta as principais abordagens utilizadas, fontes de conhecimento e inventários de sentidos em sistemas WSD. A Seção 2.5 apresenta a mineração de texto como tarefa alternativa para descobrir padrões morfô-semânticos frequentes em texto, além das explorações iniciais com padrões MSC^+ (Seção 2.5.2). Por fim, a Seção 2.6, detalha alguns dos principais obstáculos que tornam WSD uma tarefa complexa para determinados tipos de textos, além de apresentar ao leitor considerações que tornam a tese acessível a um público mais amplo.

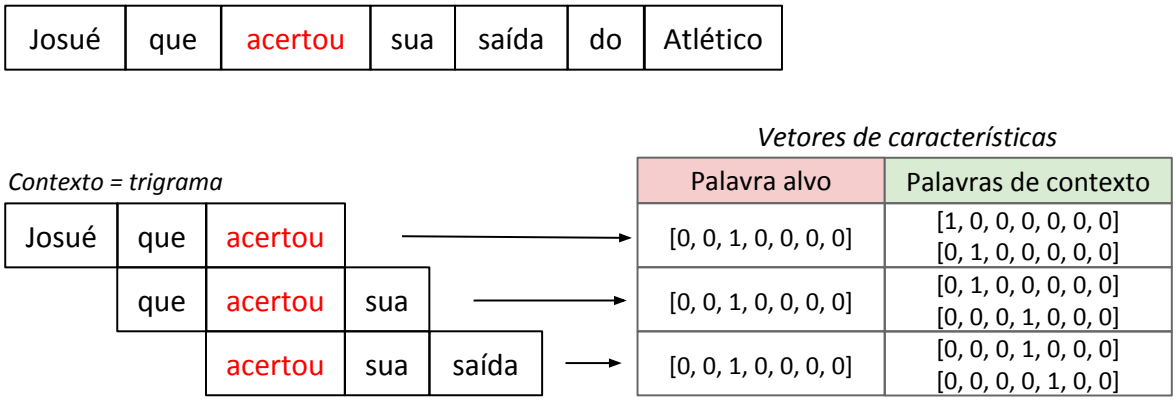
2.1 O CONTEXTO TEXTUAL

O contexto textual é qualquer informação que pode ser utilizada para caracterizar a situação das palavras (e.g., classes gramaticais, expressões multipalavras, entidades como pessoa, lugar, objeto físico ou computacional) (ABOWD *et al.*, 1999). Em geral, na tarefa de WSD o contexto de uma palavra é representado por uma sequência de palavras de tamanho $\pm n$ (e.g., *n* palavras da mesma sentença, frase, parágrafo à direita e/ou esquerda da palavra alvo) (LUND; BURGESS, 1996).

O contexto representado na forma de uma sequência de palavras de tamanho limitado, denominado **janela de palavras** ou **janela de contexto** (NAVIGLI, 2009), gera padrões estatísticos que podem ser mapeados por meio da representação das palavras no Modelo Espaço Vetorial (SEBASTIANI, 2002), de forma a explorar propriedades específicas para expressar o significado do texto (TURNEY; PANTEL, 2010). A Figura 3 ilustra um exemplo de representação do contexto por meio de vetores a partir do vetor categórico para a sentença *"Josué, que acertou sua saída do Atlético."*, com a palavra alvo $w = \text{“}acertou''$. Os vetores de características possuem sete características locais (1 na posição correspondente à palavra) que representam o contexto de tamanho trigrama acerca da palavra alvo, codificado na forma *one-hot*.

Janelas de contexto combinam diferentes hipóteses que permitem interpretar o significado de textos como uma distribuição de características específicas baseadas em modelos composicionais distribuídos (TURNEY; PANTEL, 2010; ANTUNES, 2018). Por exemplo, a (1) **Hipótese distribucional** na linguística é a de que as palavras que ocorrem em contextos similares tendem a possuir significados similares (HARRIS, 1954); a (2) **Hipótese de Bag-of-Words (BoW)** - as frequências de palavras nos textos

Figura 3 – Representação do contexto por meio de vetores.



tendem a indicar algum aspecto de seu significado (SALTON; WONG; YANG, 1975); a (3) **Hipótese distribucional estendida** - os padrões que coocorrem com pares de palavras similares tendem a possuir significados similares (LIN; PANTEL, 2001b).

Normalmente, as técnicas tradicionais de representação do contexto utilizam apenas a frequência de informações léxicas e/ou sintáticas do texto para correlacionar palavras a sentidos (e.g., coocorrência dos termos, BoW). Apesar de considerar atributos do texto como independentes, ignorando informações semânticas e contextuais, abordagens de representação do contexto inspiradas na técnica BoW são efetivas em muitas aplicações de WSD (LESK, 1986; GAONA; GELBUKH; BANDYOPADHYAY, 2009; MILLER, T. *et al.*, 2012).

As abordagens de representação do contexto baseadas em matrizes de coocorrência e decomposição de valores singulares (e.g., *Latent Semantic Analysis (LSA), Latent Semantic Indexing (LSI)* permitem correlacionar palavras com significado latente em uma coleção de texto (LANDAUER; FOLTZ; LAHAM, 1998). As técnicas LSA e LSI costumam ser mais flexíveis do que as técnicas de coocorrência dos termos e BoW, mas em virtude da multilinearidade das soluções encontradas, seus resultados podem ser difíceis de interpretar ou abstratos, sendo justificados no nível matemático, mas não interpretável em linguagem natural (COECKE; SADRZADEH; CLARK, 2010).

As abordagens de representação do contexto inspiradas em redes neurais (e.g., *word embeddings treinados com deep learning*) permitem mapear palavras individuais para vetores únicos em um espaço vetorial pré-definido e de tamanho fixo (MIKOLOV *et al.*, 2013; VIAL; LECOUTEUX; SCHWAB, 2017). Embora essas abordagens sejam mais efetivas em domínios específicos e computacionalmente mais custosas, técnicas como *word embeddings* possibilitam agregar um grande volume de informações semânticas e contextuais que geralmente são perdidas em outras técnicas de representação. Por esse motivo, em alguns casos essas técnicas superam a performance de técnicas tradicionais (ANTUNES, 2018).

As características do contexto de palavras também podem ser representadas por meio de árvores ou grafos (AGIRRE; SOROA, 2009; NÓBREGA; PARDO, 2014; SANTOS; GONÇALO OLIVEIRA, 2015). O contexto na forma de vetores é uma representação pouco estruturada, frequentemente usado em abordagens supervisionadas para WSD. A forma de grafos é uma representação estruturada, aplicada principalmente em abordagens não supervisionadas e baseadas em conhecimento (NAVIGLI, 2009).

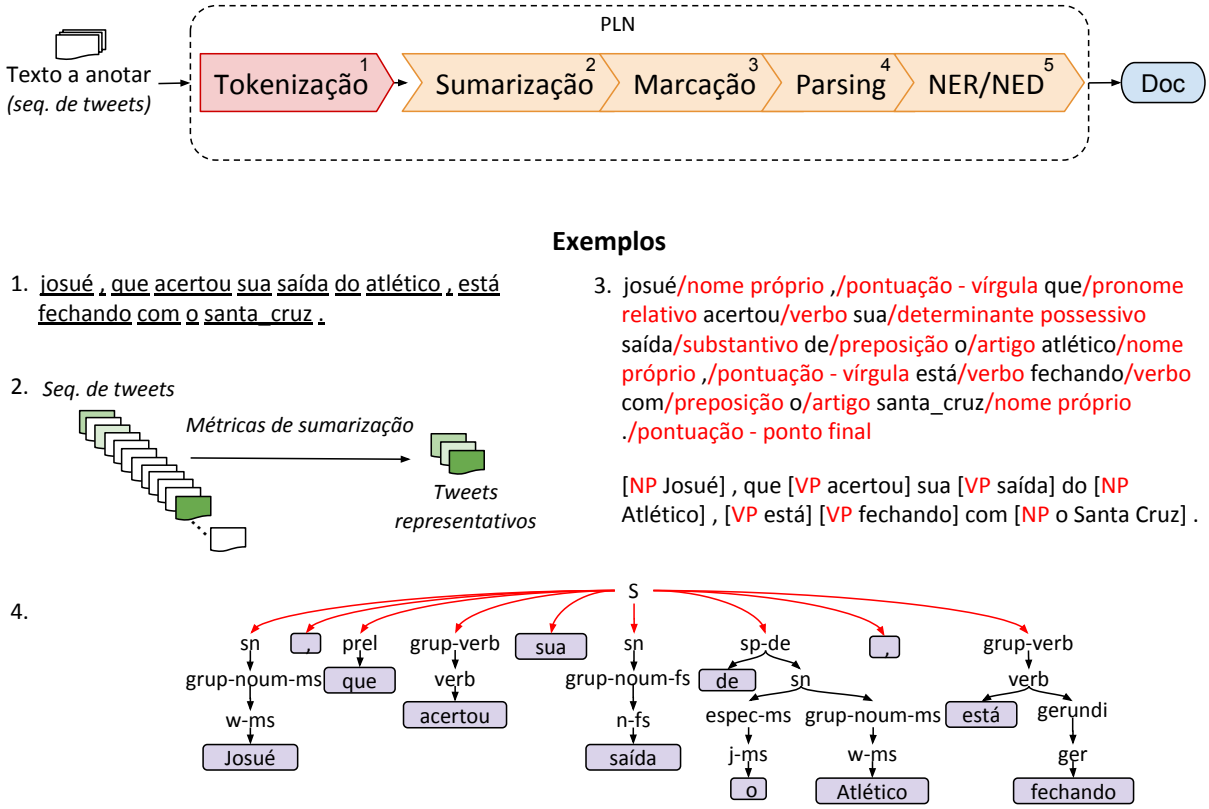
2.1.1 Características do contexto textual

A maneira tradicional de representação do contexto consiste em dividir o texto em unidades menores de informação (*tokens*) e aplicar técnicas de extração ou seleção de termos para identificar características textuais relevantes (SCHÜTZE, 1998). As características textuais utilizadas para representar o contexto são originadas das (mas não limitada a) informações de natureza léxica, sintática e/ou semântica resultantes de etapas de pré-processamento e processamento do texto. A Figura 4 ilustra um pipeline com tais etapas e exemplos.

Tokenização. Tarefa que busca dividir uma *string* (sequência de caracteres) em unidades chamadas *tokens* (palavras, palavras compostas, multi-palavras, pontuação, símbolos) (BIRD; KLEIN; LOPER, 2009). A Figura 4, exemplo 1, ilustra os *tokens* de um texto de entrada marcados com sublinhado. Derczynski, Ritter *et al.* (2013) aplica normalizações através de expressões regulares para identificar *emoticons*, nomes de usuários, *hashtags* e *URLs* em dados do Twitter, em inglês, obtendo medida-f de 96%. Porém, em postagens do Twitter, em língua portuguesa, o mesmo tokenizador não foi capaz de obter igual desempenho (SORATO; GOULARTE *et al.*, 2016). Estes novos desafios fomentam as pesquisas na área de tokenização, melhorando os métodos que foram padrões por décadas, como a tokenização PTB (*Penn Treebank*) (MARCUS; MARCINKIEWICZ; SANTORINI, 1993), além de melhorar o desempenho de outras tarefas, tais como lematização, *Stemming*, *PoS Tagging* (FOKKENS *et al.*, 2013).

Sumarização de tweets. A sumarização de *tweets* é uma tarefa que surgiu para resolver o problema de grande volume de usuários e postagens do Twitter. Consiste na extração de um grupo de *tweets* significativos de um conjunto de *tweets* (REN, 2016). Como na sumarização multidocumento (CAO *et al.*, 2016; SAGGION; POIBEAU, 2013), cada *tweet* representa um documento, e o sumário produzido prioriza a qualidade dos *tweets* selecionados, incluindo noções como representatividade e diversidade (Figura 4, exemplo 2). A maior parte dos trabalhos na literatura apresentam a sumarização de *tweets* como uma tarefa usada na detecção de tópicos e eventos, mas que pode ser explorada para outro propósito.

Figura 4 – Etapas de pré-processamento e processamento do texto.



Alguns trabalhos sumarizam *tweets* por meio de grafos (SHARIFI; HUTTON; KALITA, 2010; XU *et al.*, 2013; SCHINAS *et al.*, 2016; HE *et al.*, 2017), frequência dos termos (TAKAMURA; YOKONO; OKUMURA, 2011), probabilidade condicional (REN *et al.*, 2013), modelos bayesianos de aprendizagem (CHANG, Y. *et al.*, 2016). Outros trabalhos usam técnicas mais sofisticadas, como em Cao *et al.* (2016) que utilizam *hashtags* e *hyper-links* para gerar *clusters* sobre diferentes tópicos, sumarizando *tweets* com uma técnica de programação linear inteira.

Marcação. A tarefa *Part-of-Speech (PoS) Tagging*, *PoS Tagging*, tem o objetivo de anotar cada palavra em um texto com a sua respectiva classe morfológica (indicada por uma *tag*), com base nas definições e nos contextos dos componentes do texto (BRILL, 2000). O primeiro texto na Figura 4, exemplo 3, ilustra cada *token* anotado com uma descrição de sua classe morfológica (e.g., *"josué"* é um nome próprio). As *tags* de *PoS* ajudam na exploração de informações contidas nas palavras e indicam explicitamente algumas das estruturas inerentes ao idi-

oma (WALLACH, 2004). A técnica de *text chunking* consiste em dividir um texto em grupos de palavras sintaticamente relacionadas, e sem sobreposição (uma palavra é exclusiva de uma frase ou *chunk*) (TJONG KIM SANG; BUCHHOLZ, 2000; ATTARDI; DELLORLETTA, 2008). A técnica usa as etiquetas de um analisador morfológico para dividir e marcar o texto em frases nominais (NP) ou verbais (VP). O segundo texto na Figura 4, exemplo 3, ilustra o resultado do processo de *chunking* onde a frase *"o Santa Cruz"* forma um *chunk* ou frase nominal (uma unidade textual de *tokens* de palavras adjacentes). O processo de *Chunking* contribui com uma base de informação que permite níveis adicionais de análise e a possibilidade de adiar decisões complexas (RAMSHAW; MARCUS, 1999). Além disso, evita os problemas de *parsing* completo, como o tempo de execução alto e a baixa acurácia. Alguns dos vários métodos para aplicar *text chunking* são baseados em: máquina de vetores de suporte (KUDO; MATSUMOTO, 2001), aprendizado baseado em transformação (RAMSHAW; MARCUS, 1999), reconhecedores-NP (KUPIEC, 1993; VOUTILAINEN, 1995), estatística (CHURCH, 1988), entre outros.

Parsing. A análise sintática superficial, *Shallow parsing*, consiste na identificação dos elementos constituintes de uma frase (e.g, substantivos, verbos, adjetivos, etc.) e, em seguida, os vincula a unidades de ordem superior com significado gramatical discreto (e.g., grupos de substantivos, verbos, adjetivos, etc.) (ABNEY, 1991). A Figura 4 ilustra os elementos constituintes de um texto e suas vinculações. Como resultado da tarefa, por exemplo, os verbos *"está"* e *"fechando"* fazem parte do mesmo grupo verbal. Assim como *Chunking*, *Shallow parsing* não requer a informação sintática completa de frases, como ocorre no *parsing* completo. Isso resulta em tarefas de extração e recuperação de informação, onde os dados mais relevantes podem ser grupos de frases substantivas e verbais ou relações sintático-semânticas (e.g., relações com o verbo principal, como o sujeito e o objeto) (HAMMERTON *et al.*, 2002; ZELENKO; AONE; RICHARDELLA, 2003).

NER/NED. O processo de reconhecimento de entidades nomeadas ou *Named Entity Recognition (NER)* consiste em identificar entidades nomeadas no texto e classificar cada referência com um tipo correspondente (e.g., pessoas, lugares, organizações, filmes, produtos) (TJONG KIM SANG; DE MEULDER, 2003; NOUVEL; EHRMANN; ROSSET, 2016). A Desambiguação de Entidades Nomeadas ou *Named Entity Disambiguation (NED)* consiste em vincular as menções de entidades no texto com a sua entidade correspondente em bases de conhecimento (SHEN; WANG; HAN, 2015). A Figura 4 ilustra o resultado do processo de NER/NED com o reconhecimento e ligação de entidades nomeadas a uma fonte de dados na Web (Wikipédia). O NER/NED é explorado em textos de várias

áreas, como medicina (KIM; LU; WILBUR, 2015), biologia, química (ELTYEB; SALIM, 2014), e com destaque em mídias sociais (RITTER; CLARK; ETZIONI *et al.*, 2011; JUNG, 2012; DAS; BURMAN *et al.*, 2013; BELAININE; FONSECA; SADAT, 2016). A tarefa de NER/NED não é algo simples, pois a definição de uma entidade mencionada pode não existir na base de conhecimento, ou ainda, deve considerar a ambiguidade/sinonímia[1] de entidades (USBECK *et al.*, 2014). O desempenho de métodos de NER/NED requer atenção, geralmente porque a ligação de entidades é realizada a bases na Web, cuja a quantidade de dados é maciça (LIN; ETZIONI *et al.*, 2012).

Características do contexto representadas por técnicas baseadas no Modelo Espaço Vetorial podem comprometer o desempenho de algoritmos de AM em função da alta esparsidade (proporção de zeros) e dimensionalidade (número de características) dos dados (JURAFSKY; MARTIN, 2009). Uma possibilidade para atenuar os problemas de alta esparsidade e dimensionalidade dos dados, e auxiliar na melhora da performance de algoritmos na tarefa WSD, é agregar informações (e.g., léxica, sintática, e principalmente semântica) ao contexto (MIKOLOV *et al.*, 2013).

2.2 ENRIQUECIMENTO DO CONTEXTO

O enriquecimento do contexto é um processo que adiciona ao texto informações relacionadas à forma dos termos (**enriquecimento léxico**), à constituição das sentenças (**enriquecimento sintático**) ou ao significado (**enriquecimento semântico**) (JURAFSKY; MARTIN, 2009; AGGARWAL; ZHAI, 2012; SUNDERMANN *et al.*, 2018). As etapas de Marcação, *Parsing* e NER/NED presentes na Figura 4, podem ser utilizadas para o enriquecimento léxico, sintático e semântico, respectivamente. No caso dos *tweets*, o enriquecimento semântico é umas das principais formas de identificar os elementos que descrevem os significados de um texto.

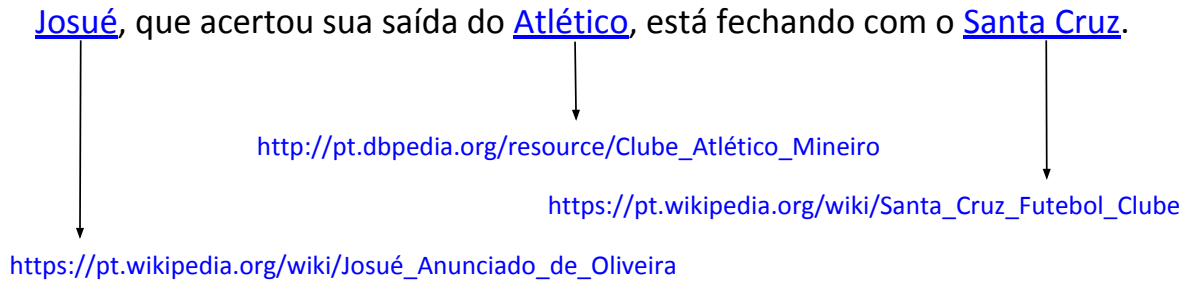
2.2.1 Enriquecimento semântico

O enriquecimento semântico ou Enriquecimento Semântico de Dados (*Semantic Enrichment of Data (SED)*) visa atribuir semântica a (porções de) dados, através de um processo que os anota com referências a recursos de informação com semântica bem definida, como conceitos ou instâncias de conceitos em ontologias, bases de conhecimento (e.g. Wikipédia) ou dados ligados (e.g. DBpedia) (BONTCHEVA; ROUT, 2014; NGOMO *et al.*, 2014; LEHMANN *et al.*, 2015), ou simplesmente palavras com

[1] **Ambiguidade** - quando entidades distintas têm nomes idênticos (e.g., *Michael Jordan*, o jogador de basquete ou *Michael Jordan*, o professor de Berkeley). **Sinonímia** - quando a entidade tem diferentes nomes (e.g. *Johann Sebastian Bach* and *J.S. Bach*).

significados precisos descritos em bases léxicas (e.g. WordNet) (DIAS-DA-SILVA; DI FELIPPO; NUNES, 2008).

Figura 5 – Típico processo de enriquecimento semântico.



A Figura 5 ilustra um texto de entrada e o mesmo texto enriquecido com dados ligados. As menções *"Atlético"* e *"Santa Cruz"* são associadas às descrições dos respectivos times de futebol na DBpedia e Wikipédia. Dados anotados, como os da Figura 5, permitem distinguir claramente coisas homônimas (e.g. *Clube Atlético Mineiro*, time de futebol de Belo Horizonte, Minas Gerais) e explorar relações semânticas entre as coisas (e.g. *Cidade do Galo* é o centro de treinamento do *Clube Atlético Mineiro*). Além disso, permitem distinguir conceitos (e.g., local, produto, evento) e instâncias de conceitos (e.g., *Belo Horizonte, Cidade do Galo*), auxiliando na desambiguação de significado e na expansão semântica.

Os critérios utilizados para selecionar os *tokens* representativos do texto podem variar, por exemplo, entidades nomeadas (ABEL; HAUFF *et al.*, 2012; SCHULZ; GUCKELSBERGER; JANSSEN, 2017), relevância ou frequência dos termos (*tf-idf*) (ABEL; GAO *et al.*, 2011; VOSECKY *et al.*, 2014) e heurísticas de identificação de local/tempo (VOSECKY *et al.*, 2014; SCHULZ; GUCKELSBERGER; JANSSEN, 2017).

O SED pode ser utilizado em *tweets* como uma maneira de adicionar contexto ao conteúdo do texto por meio de informações externas (ABEL; GAO *et al.*, 2011). De acordo com Romero (2017), diferentes técnicas e elementos podem ser explorados para enriquecer o contexto de *tweets*, inclusive de maneira híbrida:

1. **Documentos externos**: o contexto é capturado do conteúdo de documentos da Web que são identificados nos *tweets* por meio de *Uniform Resource Locators - URL* (ABEL; GAO *et al.*, 2011; VOSECKY *et al.*, 2014). Também podem-se selecionar palavras relevantes no texto para acessar, por exemplo, páginas relacionadas da Wikipédia (GENC; SAKAMOTO; NICKERSON, 2011), assim como o conteúdo da WordNet (BELAININE; FONSECA; SADAT, 2016). A dificuldade desta abordagem refere-se a definição de qual conteúdo extrair e quais dados são importantes para a tarefa de WSD;

2. **Web semântica**: o contexto é capturado através do mapeamento entre elementos extraídos dos *tweets* e recursos disponíveis na nuvem, tais como *Linked Open Data (LOD)*[2] e *Linguistic Linked Open Data (LLOD)*[3]. Esses recursos contêm propriedades semânticas descritivas sobre diferentes domínios (e.g., conteúdo gerado pelo usuário, mídias, domínios cruzados) que ajudam a generalizar o conteúdo de *tweets* (ABEL; HAUFF *et al.*, 2012; SCHULZ; GUCKELSBERGER; JANSSEN, 2017). O desafio desta abordagem está na falta de diretrizes para determinar qual fonte de conhecimento ou propriedades semânticas utilizar (e.g., *rdf:type*, *dct:subject*);

3. **Ferramentas de NER/NED**: ferramentas, tais como a Open Calais, Alchemy, DBpedia Spotlight e Babelfy são conectadas indiretamente a bases de conhecimento que são usadas para identificar as entidades nomeadas em *tweets* (ABEL; GAO *et al.*, 2011; VOSECKY *et al.*, 2014). As entidades nomeadas nas bases de conhecimento possuem informações categóricas das quais se capta o contexto. O principal desafio desta abordagem é captar o contexto relevante presente nas bases de conhecimento com precisão.

O enriquecimento semântico de elementos textuais relevantes em *tweets* é importante para sistemas WSD porque aprimora a qualidade e a eficiência das representações de palavras em razão da descoberta de atributos textuais mais relevantes para cada domínio específico, além de permitir a exploração dos significados dos dados em aplicações diversas (CEWEB, 2016).

2.3 DESAMBIGUAÇÃO LEXICAL DE SENTIDO

Cada língua natural é regida por um conjunto finito de regras gramaticais que definem as sequências de léxicos permitidos e, por consequência, padroniza a comunicação entre dois ou mais indivíduos (CHOMSKY, 1959; BALLY; SECHEHAYE, 1966; FILLMORE, 1969). A aplicação dessas regras permite definir diferentes gramáticas: as **normativas**, constituídas pelas regras gramaticais que devem ser seguidas; as **descritivas**, formadas por regras da língua falada que são seguidas independentemente das gramáticas normativas; e as **implícitas** (ou internalizadas), definidas por regras que os falantes da língua dominam (POSSENTI, 1996).

O estudo da gramática segundo os sons da fala (fonética), da estrutura das palavras (morfologia), da estrutura das frases (sintaxe), do significado das palavras (semântica) e dos estilos de escrita (estilística ou discurso), possibilita analisar os diferentes aspectos das línguas tanto pela norma culta quanto de suas variações informais (POSSENTI, 1996; TRAVAGLIA, 2003). Como o foco desta tese está na exploração

[2] `http://linkeddata.org/`
[3] `http://linguistic-lod.org/llod-cloud`

de informações semânticas e contextuais é evidenciado o estudo do significado das palavras.

Uma simples palavra pode representar diferentes sentidos em textos distintos ou de diferentes domínios (FLORIN, 2002). Geralmente, uma das formas para inferir o significado mais adequado são as outras palavras usadas no texto. Para isso, deve-se considerar o conteúdo semântico original exterior ao texto e o conteúdo semântico relacional, resultante das relações específicas que os componentes léxicos, principalmente os substantivos e os verbos, mantêm entre si no interior do texto (ARRAIS, 1975). O processo de associar a cada palavra o seu sentido ou significado apropriado de acordo com um contexto particular é chamado de Desambiguação Lexical de Sentido (DLS) (JURAFSKY; MARTIN, 2009; CABEZUDO; PARDO *et al.*, 2015), do inglês *Word Sense Disambiguation - WSD* (AGIRRE; EDMONDS, 2007; NAVIGLI, 2009).

Formalmente, um texto T pode ser considerado como uma sequência de palavras w separadas em sentenças pelo uso da pontuação, $T = \{w_1, w_2, ..., w_n\}$. Uma palavra w é uma entidade que possui uma propriedade chamada senso ou sentido, definido em um inventário de sentido[4] D. O $Sentidos_D(s_i)$ é o conjunto de todos os sentidos da palavra w_i. Então, pode-se formalmente definir WSD como uma tarefa que cria um mapeamento A, de tal forma que $A(w_i, T)$ é o correto sentido da palavra w_i dentre os $Sentidos_D(w_i)$, com $A(w_i, T) \subseteq Sentidos_D(wi)$.

Os sistemas WSD usam diferentes abordagens para encontrar o sentido de palavras, mas de uma forma geral, todos funcionam extraindo informações linguísticas do contexto da palavra alvo e a comparam com os sentidos armazenados em um inventário (MILLER, 2016). O contexto pode ser identificado por meio de software padrão de PLN (e.g., segmentadores, tokenizadores, lematizadores, *part-of-speech taggers* e *parsers* sintáticos). A Figura 6 ilustra o problema de desambiguação para a palavra alvo "*polar*" em relação aos seus sete sentidos existentes na rede BabelNet.

De acordo com a formalização apresentada $T = (w_1, w_2, w_3, w_4,\ w_5, w_6, w_7)$ = (que, vontade, de, tomar, uma, polar, trincando) e
$Sentidos_D(polar)$ = $(s_1, s_2, s_3, s_4, s_5, s_6, s_7)$. O objetivo é encontrar o correto sentido de w_6, então $A(polar) \subseteq Sentidos_D(4)$, com o $Sentidos_D(4)$ sendo o subconjunto que melhor casa com o contexto T.

Para identificar as palavras no texto muitos sistemas WSD usam ferramentas de tokenização, e em seguida aplicam *taggers* para classificar as palavras morfologicamente. Tokenizadores apresentam alta precisão e *taggers* estão atualmente disponíveis para vários idiomas (MORO; CECCONI; NAVIGLI, 2014). O resultado do *tagger* para o texto na Figura 6 pode ser T = (que/PRONOME, vontade/NOME, de/PREPOSIÇÃO, tomar/VERBO, uma/DETERMINANTE, polar/NOME,

[4] Um inventário de sentido é uma coleção de sentidos especificamente organizados para a tarefa WSD e outras aplicações em PLN (MILLER, 2016).

Figura 6 – Típico problema de desambiguação.

s_1: (nome) esta é uma lista de personagens da série de jogos.
s_2: (nome) o Polar Satellite é um satélite lançado pela NASA.
s_3: (nome) são uma classe de variáveis cataclísmicas magnéticas.
s_4: (nome) Polar Export é uma marca de cerveja pilsen brasileira.
s_5: (nome) cola da empresa brasileira Compactor.
s_6: (adjetivo) próximo ao polo geográfico (tradução).
s_7: (adjetivo) extremamente frio.

trincando/VERBO). Neste contexto, a palavra alvo "*polar*" é usada como um nome (substantivo), e somente esta informação é suficiente para descartar os sentidos s_6 e s_7 porque estes se referem a adjetivos. Para determinar qual dentre os sentidos restantes contém o significado da palavra alvo, pode-se enriquecer o texto com semântica e comparar o contexto do texto com o contexto de cada sentido. O desambiguador pode então selecionar como o único sentido correto: $A(polar) = s_4$.

A tarefa WSD pode consistir apenas na identificação da classe semântica das palavras (granularidade maior), ou na captura do significado preciso (granularidade fina). O significado preciso proporciona um conhecimento mais detalhado sobre o texto, mas é mais difícil de se capturar. De maneira geral, os sistemas WSD alcançam cerca de $60 - 70\%$ de precisão para os sentidos de granularidade fina, e $80 - 90\%$ para os mais grosseiros (MORO; RAGANATO; NAVIGLI, 2014; MORO; NAVIGLI, 2015).

2.4 SISTEMAS WSD

Os sistemas WSD, independentemente da abordagem, usam o conhecimento presente no contexto e em recursos externos (AGIRRE; EDMONDS, 2007) para resolver o problema de ambiguidade. Normalmente, os recursos externos são bases de conhecimento lexical que armazenam informações produzidas independente de contexto (AMSLER, 1984). Qualquer Base de Conhecimento Lexical (BCL) que armazena conceitos e definições pode ser convertida em um inventário de sentidos predefinido. Algumas BCLs candidatas são: dicionário, dicionário readaptável de máquina, tesauro, corpora, ontologias, redes semânticas.

2.4.1 Fontes de conhecimento

As BCLs podem ser altamente estruturadas, como bancos de dados, redes semânticas e ontologias, tais como: DBpedia (LEHMANN *et al.*, 2015), YAGO (MAHDISOLTANI; BIEGA; SUCHANEK, F., 2014), BabelNet (NAVIGLI; PONZETTO, 2010), WordNet (MILLER, G. A. *et al.*, 1990), FrameNet (BAKER; FILLMORE; LOWE, 1998), Freebase (LANIADO; MIKA, 2010). Outras são pouco estruturadas, como documentos em prosa ou listas de palavras, dicionários de máquina readaptáveis, tesauros e corpora (e.g., Wikipédia[5], Wiktionary[6], PropBank[7]) (KINGSBURY; PALMER, 2002; BARONI *et al.*, 2009; MEYER, 2013).

Algumas bases de conhecimento estruturadas armazenam informações cruzadas de outras fontes de dados (e.g., YAGO[8] conhece mais de 17 milhões de entidades nomeadas, derivadas da Wikipédia, WordNet, WikiData[9], GeoNames[10] e outras fontes de dados). Para a tarefa de WSD, as bases de conhecimento mais utilizadas são a WordNet e a BabelNet (CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016).

A **WordNet** é formada por conjuntos (*synsets*) de unidades lexicais sinônimas que expressam um único significado (MILLER, G. A. *et al.*, 1990). Cada *synset* possui as seguintes informações: (1) chave, que é a palavra que melhor representa o *synset*; (2) conjunto de palavras sinônimas, agrupadas também pela classe gramatical; (3) glosa, que é uma descrição informal do sentido do *synset*; e (4) exemplos de uso, compostos por sentenças que empregam palavras do *synset*.

Os *synsets* são organizados segundo classes lexicais (substantivos, verbos, adjetivos e advérbios). Os *synsets* em cada um desses grupos são relacionados por relações semântico-conceituais, como hiperonímia/hiponímia (generalização/especialização) e holonímia/ meronímia (todo/parte, e.g., agregação) para substantivos, troponímia (detalhamento de maneiras como se fazem coisas, tais como comunicar, falar) e implicação (*entailment* em inglês, como por exemplo, comprar implica em pagar).

A **VerbNet**, construída a partir da WordNet, agrupa verbos de maneira a respeitar as relações de significado, assim como o comportamento sintático (SCARTON; DURAN; ALUÍSIO, 2014; SCHULER, 2005). Ambas permitem distinguir, com rigor, definições e relações lexicais e semânticas. Porém, como a língua permite a pluralidade no uso de itens lexicais, há casos em que uma palavra ou expressão adquire um novo sentido, o qual não é possível captar por meras relações entre *synsets*.

A **BabelNe**t é uma rede semântica de conhecimento multilingue de ampla co-

[5] `https://www.wikipedia.org/`
[6] `https://www.wiktionary.org/`
[7] `http://propbank.github.io/`
[8] `https://github.com/yago-naga/yago3`
[9] `https://www.wikidata.org/wiki/Wikidata:Main_Page`
[10] `http://www.geonames.org/`

bertura que originou-se do alinhamento automático da Wikipédia com a WordNet (NAVIGLI; PONZETTO, 2010). Na versão 3.0, a BabelNet integra recursos como Wiktionary, Freebase e dados ligados conectando conceitos e entidades nomeadas em uma grande rede de relações semânticas, alinhados semiautomaticamente em diferentes idiomas (EHRMANN *et al.*, 2014; PANCHENKO, 2016). Adotando uma estrutura semelhante a da WordNet, cada *babel synset* representa um dado significado e contém todos os sinónimos em vários idiomas, chamados de *babel senses*. Os *babel synsets* são facilmente interpretáveis, pois apresentam definições de sentido explícito, exemplos de uso selecionados manualmente e complementados por recursos adicionais, como imagens que ilustram o sentido, relações taxonômicas e informação do domínio (PANCHENKO, 2016)

2.4.2 Abordagens clássicas de WSD

Segundo Miller (2016), as principais técnicas de WSD são categorizadas de acordo com o tipo de fonte de conhecimento que utilizam.

2.4.2.1 Baseada em conhecimento

A abordagem WSD baseada em conhecimento depende exclusivamente das informações de contexto da palavra alvo, obtidas de corpora ou recursos léxico-semânticos pré-construídos (MIHALCEA, R., 2007; KILGARRIFF *et al.*, 2014). As técnicas baseadas em conhecimento (e.g., seleção de preferências, sobreposição de sentidos, relação semântica, conectividade de grafos) são amplamente aplicáveis, porém tendem a ser menos precisas em relação às abordagens supervisionadas e não supervisionadas (MORO; RAGANATO; NAVIGLI, 2014; MORO; NAVIGLI, 2015).

Seleção de preferências: alguns dos primeiros sistemas WSD basearam-se nas restrições semânticas sobre combinações de palavras (NG; LEE, 1996; AGIRRE; MARTINEZ, 2001). Por exemplo, para o verbo *"tocar"*, com o sentido de *"tocar um instrumento musical"*, espera-se que o sujeito seja uma entidade do mundo real (e.g., pessoa, cantor) e o objeto um instrumento musical (e.g., violão, piano). Assim, se nenhum destes elementos estiverem presentes no contexto de uma ocorrência do verbo *"tocar"*, então esse sentido pode ser descartado. Embora a seleção de preferências seja uma técnica intuitiva, o conhecimento sobre classes e conceitos pode ser aprendido de maneira automática a partir de corpora (AGIRRE; MARTINEZ, 2001). Historicamente, a técnica de seleção de preferências não tem alcançado alta precisão quando comparada às outras técnicas.

Sobreposição de sentidos: a técnica de sobreposição de sentidos possui uma variedade de algoritmos que se originaram a partir do algoritmo Lesk (LESK, 1986). Algoritmos do tipo Lesk funcionam comparando cada sentido da palavra alvo, presente na base de conhecimento, com os rótulos da sua definição de contexto (e.g., *BoW*),

ou com o contexto obtido das palavras próximas da palavra alvo (NÓBREGA, 2013). Para obter o contexto das palavras, a técnica de sobreposição se beneficia de tarefas de PLN, tais como tokenização, lematização e *PoS Tagging*. O sentido correto da palavra alvo é o que apresentar a maior sobreposição de palavras. Formalmente, considerando $D(s) = (w_1, w_2, \ldots, w_n)$ a definição do sentido s, então a medida Lesk entre dois sentidos $Lesk(s_1, s_2)$ é $Lesk(s_1, s_2) = |D(s_1) \cap D(s_2)|$ (VIAL; LECOUTEUX; SCHWAB, 2017). Apesar de simples, o algoritmo Lesk e suas extensões apresentam um bom desempenho na tarefa de WSD.

Relação semântica: as bases de conhecimento altamente estruturadas permitiram o surgimento de várias medidas de similaridade semântica baseadas em grafos (SINHA; MIHALCEA, 2007; HENRICH, 2015). Por exemplo, Leacock e Chodorow (1998) definem uma medida de similaridade por meio do comprimento do caminho mais curto entre dois sentidos, normalizada conforme a profundidade da taxonomia. As medidas de similaridade entre sentidos definem uma função $sim(s_1, s_2)$ que retorna um valor entre 0 e 1 (0 significa total dissimilaridade e 1 total similaridade). A equação 2 define uma medida de similaridade entre sentidos. Considerando os sentidos candidatos da palavra alvo, realiza a computação da similaridade com o sentido de cada palavra do contexto e soma o valor máximo. O sentido candidato da palavra alvo é o que apresentar a maior similaridade com as palavras de contexto.

$$D(i) = arg \max_{s \in I(w_i)} \sum_{w_j \in T : w_j \neq w_i} \max_{\acute{s} \in I(w_j)} sim(s, \acute{s}) \tag{1}$$

Conectividade de grafos: várias abordagens WSD baseadas em grafos não mensuram diretamente a semelhança entre pares de sentidos. Em vez disso, criam um grafo de contexto para verificar quais sentidos são mais relevantes por meio de métricas de centralidade de grafos (MIHALCEA, R., 2005; SINHA; MIHALCEA, 2007). Os vértices do grafo representam todos os sentidos possíveis das palavras em um contexto e as arestas, as relações semânticas entre os sentidos. Um algoritmo como o PageRank (PAGE *et al.*, 1999) pode ser usado para determinar os vértices mais relevantes em uma rede de conexões e selecionar o sentido com maior pontuação de centralidade. O desempenho de sistemas WSD com base na conectividade de grafos pode, em alguns casos, ser superior ao de sistemas supervisionados (NAVIGLI; LAPATA, 2010; PONZETTO; NAVIGLI, 2010).

2.4.2.2 Supervisionada

As abordagens supervisionadas para WSD usam técnicas de aprendizado de máquina e necessitam de dados manualmente anotados para realizar o treinamento de um modelo (NAVIGLI, 2009). Nesta abordagem, WSD é considerado como um problema de classificação, onde uma classe (sentido da palavra) deve ser apreendida

para cada nova observação (palavra alvo). A previsão do sentido é realizada por um algoritmo classificador, previamente treinado a partir de um conjunto de dados rotulados para atribuir classes conhecidas e corretas.

As características linguísticas usadas para treinar um classificador (e.g. *tokens*, classe gramatical, correlação sintática) em WSD, assim como o tipo de classificador variam de acordo com a implementação. De acordo com Navigli (2009), os classificadores que se destacam usam modelos probabilísticos (algoritmo *Naive Bayes*) (TURNEY, 2004; CAI; LEE; TEH, 2007) e neurais (HUANG *et al.*, 2012), vizinho mais próximo (kNN), classificadores de máxima entropia, máquinas de vetor de suporte (SVMs) (ZHONG; NG, 2010), listas de decisão, tabelas e árvores.

A alta precisão é um dos motivos que tem chamado a atenção para sistemas de WSD supervisionados. Porém, os classificadores geralmente não podem desambiguar palavras que não estão presentes nos dados de treinamento, ou então, classificar palavras conhecidas com sentidos não vistos previamente. Além disso, a construção de exemplos de treinamento anotados manualmente é um processo oneroso e caro. O uso do texto em língua inglesa é predominante na maioria dos trabalhos que se enquadram nesta abordagem. Para outras línguas, os dados de treinamento são inexistentes ou escassos.

2.4.2.3 Não supervisionada

As abordagens não supervisionadas para WSD aplicam o aprendizado de máquina, mas não fazem uso de dados com sentidos manualmente anotados (MILLER, 2016). Os sistemas WSD não supervisionados aprendem a agrupar ocorrências de palavras (e.g., clusters) de acordo com os seus significados (CHEN; LIU; SUN, 2014).

Ao contrário das técnicas baseadas em conhecimento e supervisionadas, abordagens não supervisionadas para WSD não aplicam rótulos de um inventário de sentido predefinido, mas induzem sentidos distintos para palavras (*Word Sense induction (WSI)*).

O principal benefício das abordagens não supervisionadas é a sua capacidade de fazer distinções de sentido de ampla cobertura e alta precisão, sem a necessidade de dados de treinamento manualmente anotados. Contudo, podem não descobrir *clusters* equivalentes aos sentidos tradicionais em uma base de conhecimento e, por esse motivo, sua avaliação é geralmente mais difícil. Segundo (NAVIGLI, 2009), as principais abordagens para WSD não supervisionado usam métodos com base no agrupamento de contexto ou palavras e grafos de coocorrência.

2.5 MINERAÇÃO DE TEXTO

A mineração de texto pode ser considerada como uma tarefa geral da mineração de dados em que se analisa texto para obter alguma informação (HEARST, 1999). A informação na forma de linguagem natural é pouco explorada devido à sua codificação não categorizada. Em virtude deste fato, a mineração de texto visa descobrir padrões em coleções de texto que representam informações não visíveis para computadores ou pessoas.

O texto natural presente nas mídias online (e.g., Facebook, Twitter, sites de notícias) apresenta um conteúdo rico em informação semântica, mas por ser menos organizado e legível por máquinas (não estruturado), a análise deste tipo de dado encontra problemas com a ambiguidade que potencializa informações imprecisas nas sentenças das coleções de textos (WEISS *et al.*, 2010). A mineração de texto aplicada à análise do sentido e conteúdo pode aumentar a precisão e cobertura de métodos de extração de informação que necessitam de informações precisas em relação ao sentido/interpretação de palavras (e.g., eventos, pessoas, citações, emoções, sentimentos).

2.5.1 Mineração de padrões frequentes

A mineração de padrões frequentes é uma tarefa de mineração de dados que tem o objetivo de descobrir padrões na forma de subsequências recorrentes em um conjunto de dados (e.g., sequências de números ou símbolos) por meio de critérios estatísticos, tais como: frequência de ocorrência, comprimento do padrão e suporte mínimo (AGRAWAL; SRIKANT *et al.*, 1995; FOURNIER-VIGER *et al.*, 2017).

A ideia de minerar padrões sequenciais pode ser adaptada à análise de texto. Em um texto, a ordenação das palavras ou elementos relevantes nas sentenças é importante (BASHAR *et al.*, 2017). Assim, sentenças em um texto podem ser consideradas como dados sequenciais, onde encontram-se subsequências de palavras (FOURNIER-VIGER *et al.*, 2017). As subsequências de palavras mais frequentes definem padrões que podem ser usados para induzir e desambiguar o sentido de palavras (BÉCHET *et al.*, 2012).

Por exemplo, a Tabela 4 mostra os padrões de palavras encontrados usando o algoritmo Apriori[11] nos *tweets* listados na Tabela 5. A Tabela 4 mostra apenas os padrões que ocorrem pelo menos duas vezes (suporte mínimo) nos *tweets* #1 a #5 (Tabela 5). Estes padrões indicam apenas as palavras mais recorrentes e as que aparecem juntas com maior frequência. O suporte (ou suporte absoluto) de uma sequência s_a em um documento D é definido como sendo o número de sequências que contêm s_a

[11] Projetado para encontrar grupos de itens que frequentemente aparecem juntos - problema de conjuntos de itens frequentes (Frequent Itemset Mining).

e é indicado por $sup(s_a)$. Em outras palavras, $sup(s_a) = |\{s|s \subseteq s_a \wedge s \in D\}|$ (BASHAR *et al.*, 2017).

Tabela 4 – Padrões de palavras frequentes

Id	Padrão	Suporte	#
1	$\langle\{vontade\}\rangle$	2	1, 2
2	$\langle\{tomar\}\rangle$	2	1, 2
3	$\langle\{bebendo\}\rangle$	2	1, 2
4	$\langle\{gelada\}\rangle$	4	1, 2, 3, 4
5	$\langle\{tomar, gelada\}\rangle$	2	1, 2
6	$\langle\{bebendo, gelada\}\rangle$	2	1, 2

Os padrões MSC^+ propostos nesta tese são mais elaborados do que padrões sequenciais em dois sentidos: (1) levam em consideração a ordem de ocorrência das palavras, e não apenas a sua ocorrência nos documentos; e (2) as instâncias do padrão MSC^+ são comparadas de maneiras mais sutis e abstratas do que a correspondência de palavras. Os MSCs são comparados em termos de marcação de *PoS* e classe de sentido, em vez de apenas léxicos. Eles também podem ser comparados usando medidas de similaridade morfo-sintática e/ou semântica.

2.5.2 Investigação de padrões MSC^+

Uma palavra pode ser interpretada de várias maneiras, dependendo do contexto em que ocorre – fenômeno da ambiguidade lexical (ALBANO; BENEVENTANO; BERGAMASCHI, 2014; ALAGIĆ; ŠNAJDER; PADÓ, 2018). Às vezes, as palavras apresentam uma semântica implícita, (e.g., para fazer humor, ironia ou jogo de palavras). Por exemplo, o *tweet "Devassa loira, morena ou ruiva???"*, brinca com palavras que geralmente se referem à cor do cabelo juntamente à palavra "Devassa". O significado que estas palavras assumem em função do contexto é diferente daqueles significados encontrados em um dicionário de línguas, porém considerando o padrão que aparece nos *tweets* apresentados na Tabela 5, pode-se inferir o significado de *Cerveja* à palavra "Devassa", conforme explicado a seguir.

A Tabela 5 fornece um exemplo de padrão MSC^+ que surge em textos curtos (#1 a #5) extraídos do Twitter[12]. Cada *tweet* (um por linha) possui uma instância do padrão MSC^+ que é uma sequência de MSCs distintos (palavras) com as respectivas classes morfossintáticas candidatas e sentidos correspondentes à sequência: $MSC_1\langle VER, Ingerir\rangle$, $MSC_2\langle SUB, Cerveja\rangle$, $MSC_3\langle ADJ, Qualquer\rangle$. Cada instância desse padrão é uma sequência de 3 MSCs (colunas MSC_1, MSC_2 e MSC_3) destacadas em negrito nos *tweets*. Essas sequências são semelhantes em termos de marcação de *PoS* e significados de seus respectivos MSCs. O MSC_1 é um verbo que

[12] Tweets coletados usando https://twitter.com/search-home em 21 de novembro de 2018.

Tabela 5 – Exemplo de instâncias de padrões MSC^+ de comprimento 3 (MSC_1, MSC_2, MSC_3) identificados nos tweets #1 a #5

#		MSC_1 ⟨VER[1], Ingerir⟩		MSC_2 ⟨SUB, Cerveja⟩	MSC_3 ⟨ADJ, Qualquer⟩
1	Quero	**tomar**	uma	**cerveja**	**gelada**
2	Vontade de	**tomar**	aquela	**Budweiser**	**gelada**
3	Vendo jogo e	**bebendo**	uma	**loira**	**gelada**
4		**bebendo**	uma	**devassa**	**gelada**
5	Que vontade de	**beber**	uma	**polar**	**trincando**

[1]VER - Verbo; SUB - Substantivo; ADJ - Adjetivo

se refere a *ingestão de líquido* (sentido = Ingerir). O MSC_2 é um substantivo, mas apenas para os *tweets* #1 e #2 seu sentido é corretamente desambiguado para *cerveja* e *marca de cerveja*, respectivamente, usando abordagens atuais de anotação automática. O MSC_3 é um adjetivo que pode se referir a *gelado*, *extremamente frio*, entre outras possibilidades (sentido = qualquer).

As palavras alvo, destacadas em vermelho, são aquelas cujos sentidos precisam ser resolvidos. Graças ao padrão MSC^+ estabelecido pelas sequências de MSCs dos dois primeiros *tweets*, é possível induzir ou desambiguar o sentido das palavras alvo. A sequência de MSCs dos dois primeiros *tweets* apresentam uma aderência perfeita ao padrão MSC^+ gerado, enquanto que outros *tweets* têm aderência parcial. Por exemplo, cada MSC em #3 adere ao respectivo padrão MSC^+, exceto o representado pela palavra **loira** (MSC_2), onde a classe *PoS* também é um substantivo, mas o sentido é inicialmente indeterminado. Esta aderência parcial a um padrão MSC^+ permite desambiguar a palavra alvo **loira** para o sentido de *Cerveja*.

A semântica implícita induzida pelos padrões de linguagem investigados nesta tese é bastante comum em linguagem coloquial, e particularmente nas mídias sociais. Nesses casos, os métodos atuais de anotação frequentemente falham em capturar o significado correto de certas palavras, levando a resultados com baixos níveis de precisão e cobertura (BONTCHEVA; DERCZYNSKI, 2016). Além disso, alguns critérios para desambiguar o sentido de palavras são pouco confiáveis quando aplicados a textos de mídias sociais e em meios de comunicação, como a utilização de palavras com a primeira letra em caixa alta para representar nomes próprios.

Muitas soluções para tarefas de anotação semântica amplamente usadas, como Desambiguação Lexical de Sentido (WSD), Reconhecimento/Normalização/Desambiguação de Entidades Nomeadas (NER/NEN/NED) ou Vinculação de Entidades (EL) (ZHANG *et al.*, 2010; LIU, X. *et al.*, 2012), utilizam algum tipo de inventário do sentido de palavras pré-definido, como a WordNet (NAVIGLI, 2012). No entanto, criar e atualizar constantemente grandes inventários do sentido de palavras ou bases de conhecimento (e.g., WordNet (FELLBAUM, 1998), BabelNet (NAVIGLI; PONZETTO, 2010), YAGO

(MAHDISOLTANI; BIEGA; SUCHANEK, F. M., 2013)) é uma tarefa cara e difícil. Como resultado, importantes sentidos de palavras atuais usadas nas mídias sociais, por grupos específicos, em determinadas regiões geográficas ou em domínios particulares, podem não estar armazenados em inventários (e.g. como os sentidos das palavras alvos em #3 a #5).

2.6 CONSIDERAÇÕES

Neste capítulo foram discutidos temas relacionados à análise do significado de palavras em texto, explanando conceitos básicos sobre Contexto, Enriquecimento Semântico, Desambiguação do Sentido de Palavras e Mineração de Texto, que são essenciais para a contextualização deste estudo. A abordagem proposta nesta tese para desambiguar o sentido de palavras é voltada (mas não exclusiva) a texto de mídia social. Para ajudar o leitor na compreensão dos temas apresentados, nas subseções a seguir são apresentadas algumas considerações importantes para a definição do escopo desta pesquisa.

2.6.1 Origem do problema

O problema básico das abordagens WSD origina-se dos seguintes fatos: nas últimas décadas, uma grande quantidade de informação textual tornou-se disponível em repositórios na Internet e intranets (e.g. notícias on-line e comentários, documentos governamentais, relatórios de empresas, decisões judiciais, registros médicos, pesquisas acadêmicas e mídias sociais) (PISKORSKI; YANGARBER, 2013; ZHAI; MASSUNG, 2016).

Grande parte dessa informação não tem semântica bem definida e processável por máquinas porque são documentos de texto livre. Ainda que a informação seja compreensível a seres humanos, não apresenta um formato legível para máquinas. Para que máquinas possam entender um texto livre, há necessidade de extrair informações a fim de descobrir e representar um conhecimento relevante de maneira estruturada (DEDEK, 2012). Quando a informação torna-se legível para máquinas, outro nível de informação é produzido e, a seguir, disponível aos seres humanos. Portanto, há um processo de associação entre dois níveis de informação.

Tarefas de extração de informação de texto têm o objetivo de identificar tipos de documentos, classes de entidades pré-especificadas (e.g., sintagmas nominais que denotam pessoas, grupos de pessoas, organizações, referências geográficas) (DAIBER *et al.*, 2013), conceitos em um domínio usando certos corpus e *datasets*, expressões numéricas, eventos e relações semânticas entre essas coisas. Tais tarefas permitem converter texto livre em informação estruturada e armazená-la em repositórios. A Figura 7 mostra um exemplo de um trecho de um artigo de notícias sobre uma

descoberta científica e informação estruturada derivada dele.

Figura 7 – Exemplo de extração de informação de um artigo de notícias.

Londres, Reino Unido. O desafio do balde de gelo que viralizou em 2014 conseguiu arrecadar fundos para uma descoberta científica importante: a identificação de um novo gene que contribui para a Esclerose Lateral Amiotrófica (ELA), o Nek1...

TIPO:	Ciência/Tecnologia
SUBTIPO:	Genética
LOCAL:	Londres, Reino Unido
DOENÇA:	Esclerose Lateral Amiotrófica (ELA)
GENE:	Nek1
CAMPANHA:	desafio balde do gelo
PROPÓSITO:	arrecadar fundos

No caso do artigo da Figura 7, algum domínio específico de conhecimento é necessário a fim de agregar corretamente a informação extraída em uma forma estruturada – a compreensão do fato de que o desafio do balde do gelo é uma brincadeira publicada em mídias sociais não está no texto. Para um leitor ter essa compreensão é necessário um conhecimento prévio sobre os fatos da notícia, ou então, relacioná-los a outras informações. Assim, explorar a informação contextual contida, por exemplo, em uma base de conhecimento, sobre o que é mencionado em um texto é uma das maneiras para ampliar o seu sentido e compreensão.

Devido à complexidade e à ambiguidade da linguagem natural, WSD é uma tarefa não trivial. Há muitas maneiras de expressar uma informação por meio de frases, e isso exige uma quantidade significativa de conhecimento para discernir sobre o significado da linguagem natural. Repositórios de textos, por exemplo, apresentam várias frases relacionadas a um mesmo conceito ou sentido, porém com diferentes vocábulos e construções linguísticas.

2.6.2 Mídias sociais

De uma forma simples, o texto presente nas mídias sociais (e.g., blogs, microblogs, portais de comércio eletrônico, compartilhamento de multimídia, redes sociais, plataformas de revisão de conteúdo, jogos sociais, mundos virtuais) pode ser entendido como a representação linguística das ideias de um usuário que utiliza a web para criar e compartilhar conteúdo (AICHNER; JACOB, 2015). Esse tipo de texto pode apresentar-se de tamanho e forma variados (e.g., uma única palavra, conjunto de caracteres, símbolo, imagem), e ainda assim transmitir uma mensagem de sentido completo. Porém, para que tenha sentido, alguns fatores como o contexto e o conhecimento do leitor são indispensáveis (LEU *et al.*, 2004).

O texto também pode estar combinado a recursos, mídias e ligado dentro de redes de informações complexas que exigem novas capacidades de compreender e utilizar informações em vários formatos a partir de fontes disponíveis na Internet (GILSTER, 1997). Dessa forma, palavras e expressões pertencentes ao campo lexical e ao semântico variam de sentido e de valor de acordo com a época, com a situação comunicativa, com a intenção e com o contexto em que foram empregadas.

De acordo com Ren (2016), ao contrário da mídia tradicional, os documentos de mídia social possuem características únicas em vários aspectos: (1) **Brevidade**, possuem pouco texto em relação à mídias tradicionais; (2) **Multilingual**, pessoas que falam diferentes idiomas pertencem a redes sociais que usam a mesma plataforma de comunicação (e.g., durante um evento esportivo, como os Jogos Olímpicos de 2016 - Rio 2016, as pessoas discutem sobre um determinado jogo em vários idiomas; (3) **Opiniões**, as mídias sociais são um grande repositório sobre as opiniões das pessoas, o que tem tornando a análise do sentimento, cada vez mais, importante; (4) **Disponibilidade**, os documentos de mídia social são publicados com marcações de tempo específicas, e as propriedades estatísticas do texto mudam ao longo do tempo.

O Twitter é um sistema de microblog que se tornou uma grande plataforma[13] de mídia social. O Twitter tem chamado a atenção por ser um ambiente *up-to-date*[14] sobre informações de usuários, eventos, opiniões das pessoas, tendências de negócios, etc (MEIJ; WEERKAMP; RIJKE, 2012). Por outro lado, os *tweets* podem ser agramaticais – apresentam palavras irregulares ou frases mal construídas, sem sentido aparente, pois ignoram as convenções sintáticas da língua.

2.6.3 A semântica em tweets

Semântica refere-se aos significados que podem ser atribuídos a alguma coisa (parte de um texto, imagem, som, etc.). Ela pode se apresentar de diferentes formas e sua elucidação, representação e uso têm sido pesquisados em diferentes campos de estudos[15]. Esta diversidade faz com que esse assunto seja muito discutido. Na área da Linguística não há um acordo sobre qual é a melhor teoria semântica (e.g., proposicional, não-proposicional, mentalista e não-mentalista) (SPEAKS, 2016), nem sobre quais problemas tais teorias deveriam abordar e quais dados elas deveriam analisar (FODOR, 2007).

A noção de significado de uma expressão abarca os conceitos de *sentido* e *referência* (FLORIN, 2002). A *referência* de uma expressão é um apontamento para

[13] O Twitter possui uma audiência total de 326 milhões de usuários ativos mensais e uma geração diária de centenas de milhões de pequenos textos (*tweets*) (TWITTER, 2018).

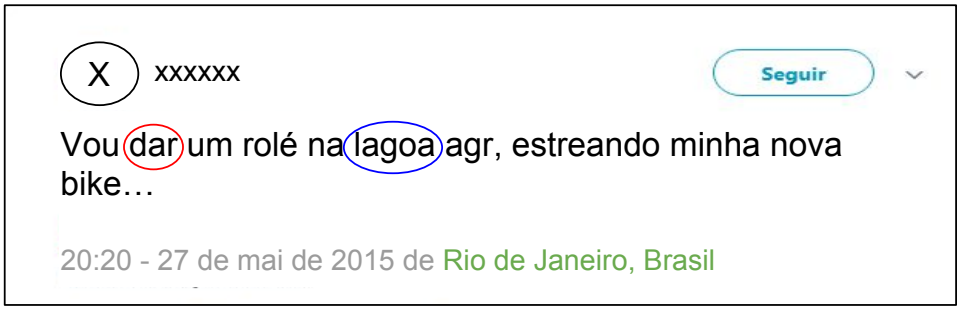
[14] Que inclui as informações mais recentes; atualizado.

[15] Os seres humanos usam diferentes tipos de semântica (e.g., léxica, sintática, discursiva) para se expressar através da linguagem (CRUSE, 2011). Já o computador precisa de um conjunto de regras sintáticas e semânticas para definir um programa. A lógica formal, por sua vez, usa semântica nas condições de verdade das proposições.

alguma coisa do mundo real (e.g., conceito, entidade, objeto, indivíduo, fato, evento), enquanto que o *sentido* é o modo como tal coisa se apresenta (HILPINEN, 2015). Para a semântica formal o significado é a relação entre uma sentença na linguagem e as coisas que ela denota do mundo real, descritas como externas à própria língua. Semântica formal implica conhecer em que condições uma sentença é verdadeira ou falsa (SPEAKS, 2016).

A Figura 8 ilustra um *post* publicado no Twitter com o texto *"Vou dar um rolé na lagoa agr, estreando minha nova bike..."*. Uma série de informações ocultas acompanham a enunciação do texto.

Figura 8 – Semântica implícita em texto de mídias sociais.



Por exemplo, se a palavra *"dar"* (circulada em vermelho) fosse tomado ao pé da letra, seria desambiguada para o sentido do verbo *"dar"*, e seu real significado não seria captado, pois depende do contexto (*"um rolé"*) para inferir o sentido de *"sair, ir, passear"*. Com mais esforço, também é possível intuir que a palavra *"lagoa"* (circulada em azul) se refere a um lugar específico, *"Lagoa Rodrigo de Freitas"*[16], na cidade do Rio de Janeiro, pois o texto do Twitter pode conter informações implícitas e de contexto temporal e georreferenciado (e.g., hora, data, local).

A semântica implícita ou latente, refere-se à relação que componentes de uma sentença assumem, indiretamente com um significado (DEHAENE *et al.*, 2006; MARTIN, 2015). Textos de mídias sociais podem adquirir novos significados pelo contexto (espaço, tempo) e pela forma como as palavras são representadas e estruturadas na composição dos textos. Pode não existir a real intenção de ocultar o sentido de uma mensagem, como ocorre na criptografia, mas alguns significados podem ficar em um estado de inacessibilidade ou latente, exigindo mais processamento semântico. Todavia, informações implícitas não são capturados pelos sistemas de WSD tradicionais, que apenas se concentram nas informações explícitas.

[16] Local onde as pessoas praticam esportes (e.g., remo, canoagem, caminhada, ciclismo) e aproveitam momentos de relaxamento. `https://pt.wikipedia.org/wiki/Lagoa_Rodrigo_de_Freitas`

2.6.4 Desafios

A análise de dados de repositórios de textos contribui para que aplicações do mundo real recuperem ou descubram um conhecimento, mas como a linguagem natural é flexível (ZOCK, 1988; BATEMAN; ZOCK, 2016) e existe uma vasta quantidade de dados, ferramentas que processam texto encontram dificuldades em tarefas de pré-processamento, desambiguação de sentido, anotação, aprendizado com os dados e no uso das informações (BONTCHEVA; DERCZYNSKI, 2016).

Diversas aplicações que usam dados de mídias sociais precisam analisar eficientemente a sintaxe e a semântica de grandes quantidades de dados gerados constantemente, cujos padrões linguísticos podem ser dinâmicos. De acordo com Morgan e Van Keulen (2014), os principais desafios para se extrair informações dos textos de mídias sociais são: (1) **Linguagem informal**, os textos publicados estão escritos em linguagem informal, incluindo erros ortográficos, falta de pontuação e maiúsculas, uso de abreviações sem padrão. Em consequência da falta de formalidade, tarefas tradicionais de processamento de texto, tais como tokenização, *POS tagging*, *Shallow parsing*, têm o seu desempenho comprometido (DERCZYNSKI; RITTER *et al.*, 2013; DERCZYNSKI; MAYNARD *et al.*, 2015; SORATO; GOULARTE *et al.*, 2016). Por exemplo, *POS taggers* alcançam 97% de precisão em textos semelhantes aos usados para o treinamento da ferramenta (principalmente textos de notícias), enquanto que conteúdos de mídias sociais a sua precisão diminui para $70 - 80\%$ (BONTCHEVA; DERCZYNSKI *et al.*, 2013). (2) **Pouco contexto**, existe um limite de texto para publicar em algumas mídias sociais, como o Twitter. Este limite obriga os usuários a usar abreviações para expressar mais informações em suas postagens. (3) **Ruído** cerca de 40% do conteúdo das mensagens no Twitter são inúteis. Em virtude desse fato, faz-se necessário usar uma etapa de pré-processamento (Filtragem) para selecionar o fluxo de postagens com o melhor conteúdo semântico. (4) **Entidades desconhecidas**, as postagens podem expressar informações sobre eventos locais, pouco significativos (e.g., festivais de rua, jogos escolares) ou recentes, portanto, as entidades envolvidas não devem estar contidas em bases de conhecimento. (5) **Contexto incerto**, é necessário lidar com a incerteza envolvida nos fatos extraídos, pois nem todas as informações disponíveis são confiáveis. As informações postadas por usuários podem ser parciais e estão sujeitas à evolução ao longo do tempo.

Apesar das dificuldades para extrair informações de postagens feitas em mídias sociais e, particularmente de *tweets* (GATTANI *et al.*, 2013), pode-se encontrar nos *posts* componentes relevantes e semanticamente ricos, tais como entidades nomeadas (menções a pessoas, locais, organizações, datas, etc.) (TJONG KIM SANG; DE MEULDER, 2003), além de palavras com classificação gramatical (substantivos, adjetivos, verbos, etc.) (JURAFSKY; MARTIN, 2009), que podem auxiliar no entendimento do que se passa com os usuários que fazem as postagens.

2.6.5 Aplicação

A tarefa de desambiguação lexical de sentido é um problema de classificação natural em Inteligência Artificial (IA) e essencial para melhorar o desempenho de várias aplicações que processam a linguagem natural (AGIRRE; EDMONDS, 2007), tais como **análise de sentimento** (DAS; ACHARJYA; PATRA, 2014), **mineração de texto** (LINOFF; BERRY, 2011), **recomendação** (PAZZANI; BILLSUS, 2007), **geração de respostas** (AL-HARBI; JUSOH; NORWAWI, 2017), **simplificadores de texto** (SIDDHARTHAN; MANDYA, 2014), **sumarizarizadores** (RUDRA *et al.*, 2016), **tradução de máquina** (JONES *et al.*, 2012) e **bases de conhecimento lexical** (MILLER, G. A. *et al.*, 1990; SCARTON, 2013; CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016)

As aplicações que processam a linguagem natural precisam utilizar diferentes aspectos da linguagem, desde unidades estruturais (e.g., fonemas, lexemas[17], frases e texto) a níveis de processamento da língua (fonologia, morfologia, sintaxe, semântica e pragmática) (MILLWARD; HAYES, 2011). O **nível fonológico** está relacionado à análise dos fonemas da língua; o **nível morfológico** relacionado aos morfemas utilizados na construção de palavras, subdividido em grupo dos afixos (prefixos, sufixos, radicais, entre outros), grupo da análise morfológica (gênero, número, grau, modos e tempos das conjugações verbais, entre outros), classes gramaticais (substantivo, verbo) e subnível léxico (formas canônicas); o **nível sintático** está relacionado ao arranjo dos lexemas em uma frase; o **nível semântico** está relacionado ao significado dos lexemas e das frases; e o **nível pragmático** está relacionado ao propósito do falante ao produzir frases.

O processo de extração e classificação de informações do conteúdo de microblogs é particularmente complexo para métodos e ferramentas de processamento de linguagem natural (JURAFSKY; MARTIN, 2009) e alternativas clássicas de processamento de texto (CAMBRIA; WHITE, 2014). Contudo, é possível estruturar um conjunto de técnicas que podem ser usadas para reconhecer e extrair informações importantes de microblogs, e assim fornecer uma abordagem para entender os possíveis significados que uma mensagem assume.

[17] É a parte de uma palavra que constitui uma unidade mínima dotada de significado lexical

3 REVISÃO SISTEMÁTICA DA LITERATURA

A identificação da semântica de palavras é particularmente útil para várias tarefas que processam texto (Seção 2.6.5). Especialmente quando o contexto do conteúdo é gerado por usuários na Web e comunicação on-line (e.g. Twitter), novos sentidos são continuamente criados por meio do uso de palavras existentes em contextos variados, que por sua vez trazem desafios adicionais à tarefa WSD. Na maioria dos casos, os inventários de sentido e sistemas de semântica lexical não estão preparados para lidar com o significado do texto em contexto pouco comum, que requer um processamento mais apurado e gradativo nos três níveis de conhecimento linguístico: **léxico**, **sintático** e **semântico** (AGGARWAL; ZHAI, 2012).

Graças ao progresso da capacidade de processamento computacional e à evolução das técnicas de PLN, os três níveis linguísticos têm sido mais amplamente explorados em WSD, com um crescente interesse no nível semântico (JURAFSKY; MARTIN, 2009; GOULARTE; SORATO *et al.*, 2019). A definição mais apurada da semântica tem um papel fundamental no significado do texto, e pode ser crucial para diferenciar o sentido de palavras que apesar de possuírem o mesmo vocabulário (i.e. conjunto das palavras que ocorrem num determinado contexto de uso), apresentam ideias diferentes ou novas. Assim, vários pesquisadores de WSD passaram a incorporar técnicas que permitem um processamento mais profundo dos textos, com o objetivo de melhorar os resultados de trabalhos.

A dependência entre texto e contexto pode ser explorada na forma de padrões ou pares de palavras que combinam os três níveis de conhecimento linguístico. Dessa forma, foi realizada uma revisão sistemática da literatura com o objetivo de identificar os estudos sobre WSD que exploram padrões linguísticos. A subseção a seguir descreve o plano de pesquisa efetuado para realizar a revisão da literatura, o processo de revisão sistemática da literatura que foi adotado e seus resultados. Os trabalhos identificados além do processo sistemático realizado complementam informações obtidas, mas não foram considerados nas estatísticas quantitativas apresentadas.

3.1 PLANO DE PESQUISA REALIZADO

Com o objetivo de investigar o interesse científico pelo estudo do significado que o texto pode assumir por meio do uso de palavras em contextos variados, durante o desenvolvimento desta tese foi realizado um levantamento bibliográfico em forma de Revisão Sistemática (RS), com o foco na tarefa de WSD. A RS é uma forma de estudo secundário que adota um método formal para a recuperação e análise crítica da literatura (KITCHENHAM; CHARTERS, 2007). A Revisão Sistemática (RS) é aplicada à análise detalhada de um volume reduzido de publicações, diferente do processo de Mapeamento Sistemático (MS), que promove uma análise exploratória de grandes

volumes de publicações. Em virtude do uso dos termos de pesquisa mais específicos, em uma RS, um número menor de trabalhos são identificados e a análise exploratória é realizada com maior profundidade (KITCHENHAM; BRERETON *et al.*, 2009; PETERSEN; VAKKALANKA; KUZNIARZ, 2015). Uma RS ajuda a pesquisadores a responder questões de pesquisa específicas, assim como identificar lacunas na pesquisa em determinado domínio de estudo, tais como: principais técnicas e algoritmos utilizados, bases de dados e *datasets*, métricas de avaliação e resultados dos trabalhos.

A RS apresentada nesta subseção foi baseada no procedimento descrito em Kitchenham, Brereton *et al.* (2009), mas com ressalva para a participação dos indivíduos envolvidos e a consideração de estudos secundários (KITCHENHAM; BRERETON *et al.*, 2009). Ainda que o plano de pesquisa tenha sido elaborado por dois pesquisadores, as fases de seleção e extração foram conduzidas por apenas um pesquisador. Com relação aos estudos secundários, como a RS realizada nesta tese tem o objetivo de identificar o maior número possível de características das produções científicas da área, mas relacionadas a uma questão de pesquisa restrita, alguns trabalhos que revisam estudos primários (e.g. *surveys*) também foram considerados. Os principais itens da RS realizada são apresentados a seguir.

Questões de pesquisa. De maneira geral, as questões de pesquisa para uma RS são questões de nível mais baixo, portanto, maior profundidade. O plano de pesquisa definido considerou uma questão principal e outras secundárias:

Principal: *Com qual frequência e como padrões linguísticos são abordados em WSD?*

Secundárias:

1. Quais são os principais domínios de aplicação?
2. Quais são os idiomas frequentemente considerados em experimentos?
3. Quais são as fontes de conhecimento/BCL comumente utilizadas?
4. Qual é o processo que prevalece - desambiguação e/ou indução?
5. Quais técnicas e algoritmos são frequentemente empregados?
6. Quais são as principais abordagens utilizadas?

Fontes de pesquisa e *string* de busca. Os trabalhos relacionados ao tema foram recuperados por meio de pesquisas nas bibliotecas e indexadores da área da computação: IEEE Xplore, Science Direct, ACM Digital Library, Web of Science e Scopus. As pesquisas foram realizadas com a seguinte *string* de busca ("word sense disambiguation" OR *"word sense induction") AND ("lexical semantics" OR semantic enrichment OR context OR classification OR tweets)*, construída a partir de termos identificados de artigos relevantes para o tema deste estudo. Em

alguns casos, a base de dados Google Scholar também foi utilizada como fonte de pesquisa, porém seus resultados não fazem parte da RS realizado.

Estratégias de busca, seleção e extração. A RS realizada consistiu em três etapas integrantes: busca, seleção e extração de informações. Os trabalhos retornados pelas **buscas** (756) passaram pela etapa de seleção manual para identificar os estudos inadequados ou irrelevantes ao tema. Os trabalhos foram selecionados pela análise do título, resumo e outras informações importantes (e.g. palavras-chave, metodologia). Os estudos que se encaixaram em ao menos um dos critérios de exclusão a seguir foram removidos (621), assim como os estudos que não atenderam aos critério de qualidade. O mesmo estudo recuperado em mais de uma fonte de pesquisa foi considerado duplicado (123), e neste caso, manteve-se apenas uma entrada por estudo.

Critérios de exclusão e qualidade:

- Publicações, tais como: resumo, resumo estendido, pôster, apresentação, editorial, ou ainda tutorial.
- Publicação sem acesso completo (e.g. hospedada em serviços inacessíveis).
- Publicação em outro idioma (não está em Inglês ou Português).
- Publicação que não se refere ao escopo do tema.

Critérios de qualidade:

- Publicação com descrição metodológica e resultados adequados.
- Publicação que evidencia os tópicos identificados nas questões de pesquisa secundária.

Finalmente, na etapa de **extração** de informações, foram analisados 135 trabalhos resultantes da etapa de seleção com o objetivo de responder às questões de pesquisa apresentadas nesta subseção. A Tabela 6 apresenta a quantificação dos estudos recuperados nas etapas de busca e seleção[1].

3.2 RESULTADO DA REVISÃO SISTEMÁTICA

A RS dos 756 estudos selecionados das fontes de pesquisa (Tabela 6) foi conduzida em dois ciclos, o primeiro em setembro de 2017 e o segundo, que corresponde a uma atualização, em fevereiro de 2019.

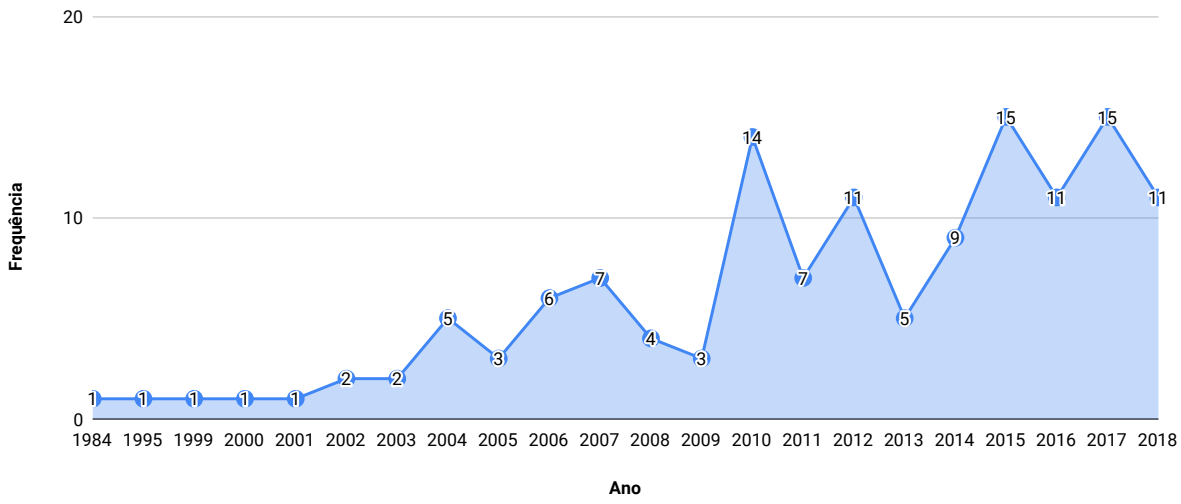
[1] Disponível em `https://github.com/fabiobif/MSC-patterns`

Tabela 6 – Estudos recuperados nas etapas de busca e seleção

Fonte de pesquisa	Resultados da busca	Resultados da seleção
IEEE Explorer	106	44
Science Direct	457	34
ACM Digital Library	16	7
Web of Science	31	2
Scopus	146	48
Total	756	135

Os resultados confirmaram alguns indícios prévios dos pesquisadores sobre o tema de pesquisa (e.g., maior número de publicações nos últimos anos) e revelaram outros pontos interessantes da área (e.g. o uso de inventários de sentido). A Figura 9 apresenta a evolução de publicações sobre o tema desde 1992 até o ano de 2018.

Figura 9 – Evolução das publicações em desambiguação de palavras.



Com base nas respostas das questões de pesquisa (Subseção 3.1), os resultados na forma de meta-análise quantitativa dos estudos selecionados são apresentados conforme os tópicos a seguir:

1. **Domínios de aplicação**: Saúde e Ciências Biológicas (40%), documento web (45%) e notícias (65%).

2. **Idioma**: Inglês (71%), italiano (25%), francês (30%),

3. **Fonte de conhecimento**: WordNets (71%) e enciclopédias livre (50%).

4. **Processo de WSD**: Desambiguação (90%) e Indução (10%).

5. **Técnica ou algoritmo**: Semântica latente, algoritmos tradicionais de AM, algoritmo de Lesk e suas variações.

6. **Abordagem**: grafos, vetores, probabilidade, Ensemble learning, Word embedding, Deep learning.

Em relação aos domínios de aplicação, que corresponde a área de pesquisa ou gênero do texto em que se executa a tarefa de desambiguação do sentido de palavras, a revisão mostrou que existe maior concentração de estudos na Saúde e Ciências Biológicas (40%), Documento web (45%) e Notícias (65%). A área da Saúde e Ciências Biológicas, com destaque à bioinformática e biomedicina, apresenta vários estudos sobre a utilização da tarefa WSI/WSD para encontrar o sentido correto de termos técnicos ou acrónimos em fontes como MEDLINE e PubMed. Contudo, os estudos que mais contribuem para a desambiguação de palavras são encontrados nos textos originados da Web e Notícias, principalmente os disponibilizados em conjuntos de dados como SenEval e Semseval.

A presença do idioma inglês é preponderante nos textos utilizados para estudos da área (71%). Depois do inglês, idiomas como chinês (25%), italiano (25%), francês (30%), espanhol (30%), alemão, português, árabe, ou mesmo bengali também são explorados. Observando a exploração dos idiomas nos estudos sobre a desambiguação de palavras, textos no idioma português foram considerados em apenas 20% dos estudos selecionados.

As WordNets e enciclopédias livre (e.g. Wikipedia) destacam-se como as fontes de conhecimento/BCL externas mais citadas nos estudos da área (71% e 50%, respectivamente). Outras fontes de conhecimento/BCL como a BabelNet (5%), SUMO (5%), OntoNotes (10%) e DBpedia Spotlight (5%) também foram citadas e exploradas nos estudos da área. Algumas destas fontes foram criadas incluindo dados de outras fontes externas. Por exemplo, a BabelNet foi criada automaticamente combinando dados da Wikipédia e WordNet e ligados à DBpedia Spotlight; YAGO combina dados da Wikipédia, WordNet e GeoNames. Este tipo de fonte de conhecimento tende a ganhar importância à medida que mais estudos sobre o tema forem publicados.

Considerando o tipo de processo, a desambiguação (90%) é predominante em relação à indução do sentido (10%). Em relação às técnicas e algoritmos, destaca-se o tratamento da semântica latente, a aplicação de algoritmos tradicionais de AM, algoritmo de Lesk e suas variações. Sobre as principais abordagens (Q1.6), destaca-se o tratamento da semântica por meio de grafos. Abordagens mais recentes como a incorporação de palavras (Word embedding) para aprendizado profundo (Deep learning) em textos é uma tendência nos estudos mais recentes na área. O aprendizado de padrões linguísticos é uma abordagem pouco frequente em WSD.

Muitas abordagens WSD/WSI são baseadas em técnicas de cluster não supervisionadas. Dentre as principais abordagens, pode-se destacar de acordo com a literatura existente: (i) agrupamento baseado em informações contextuais, (ii) abordagens baseadas em informações de palavras, (iii) técnicas baseadas em grafos e (iv)

modelos probabilísticos.

A abordagem padrão é a baseada em informações contextuais, na qual o contexto de instâncias de palavras são representados como modelo de espaço vetorial (e.g., bag-of-words) de primeira ou segunda ordem. Os vetores de contexto são obtidos por técnicas estatísticas e frequentistas como a co-ocorrência de palavras em um contexto ou por incorporação de palavras (SCHÜTZE, 1998; LI; JURAFSKY, 2015). Após a representação do contexto de palavras, os vetores de contexto gerados são agrupados em clusters de sentido (e.g., algoritmos K-means e Brown).

A segunda abordagem consiste em agrupar palavras que são semanticamente semelhantes. Normalmente, a semelhança/relacionamento entre as palavras é medida em termos de dependências sintáticas (LIN, 1998).

A terceira abordagem é baseada no agrupamento de grafos (PANCHENKO *et al.*, 2017; RODRIGUEZ *et al.*, 2016; PELEVINA *et al.*, 2017; CHANG, H.-S. *et al.*, 2018; GUTIÉRREZ; VÁZQUEZ; MONTOYO, 2017). De acordo com Di Marco e Navigli (2013), essa abordagem representa instâncias de palavras semelhantes ou relacionadas a cada palavra alvo como nós em um grafo de similaridade, que são agrupados por meio algoritmos de agrupamento de grafos (e.g., HyperLex (VÉRONIS, 2004), PageRank (AGIRRE; MARTINEZ *et al.*, 2006), Chineses Whispers (BIEMANN, 2006)).

A quarta abordagem representa sentidos na forma de modelos estatísticos, tais como Framework Bayesiano (BRODY; LAPATA, 2009), Redes Bayesianas no Aprendizado da Incorporação de palavras (WU *et al.*, 2018), ou processo hierárquico de Dirichlet (LAU *et al.*, 2012). Primeiro, para cada palavra ambígua é criado um conjunto de sentidos usando a distribuição de probabilidade sobre as palavras. Desta forma, as palavras de contexto são geradas de acordo com a distribuição de probabilidade, e diferentes sentidos podem ser obtidos, com diferentes distribuições de palavras.

As principais técnicas de avaliação para sistemas WSD são não supervisionadas (ALAGIĆ; ŠNAJDER; PADÓ, 2018; JURGENS; KLAPAFTIS, 2013; PELEVINA *et al.*, 2017) e supervisionadas (JURGENS; KLAPAFTIS, 2013; PANCHENKO *et al.*, 2017; PELEVINA *et al.*, 2017). Para a avaliação não supervisionada, os sentidos induzidos (clusters) são comparados diretamente com uma anotação referente a padrão ouro usando medidas como F-score pareado (ARTILES; AMIGÓ; GONZALO, 2009), V-measure (ROSENBERG; HIRSCHBERG, 2007) e Fuzzy NMI (JURGENS; KLAPAFTIS, 2013). A medida F-score calcula a média harmônica da precisão e cobertura, considerando como positivos verdadeiros todos os pares de instâncias agrupados em conjunto nos clusters de sentido induzido e de padrão ouro (ALAGIĆ; ŠNAJDER; PADÓ, 2018). O V-measure calcula a homogeneidade e integridade dos clusters. A homogeneidade tem por objetivo verificar se todos os seus clusters contêm apenas pontos de dados que são membros de uma única classe, enquanto a integridade representa se todos os pontos de dados que são membros de uma determinada classe são elementos do

mesmo cluster (ROSENBERG; HIRSCHBERG, 2007).

No segundo tipo de avaliação, supervisionado, os sentidos induzidos são mapeados para sentidos do padrão-ouro usando uma heurística de mapeamento, onde as medidas clássicas de precisão e cobertura são usadas para determinar a qualidade do sistema WSD resultante. Normalmente, os sistemas de linha de base (*baselines*) são usados para comparar os resultados dos sistemas WSD, como o sentido mais frequente (Most Frequent Sense - MFS) para cada palavra de um conjunto de dados (JURGENS; KLAPAFTIS, 2013; MORO; NAVIGLI, 2015; ALAGIĆ; ŠNAJDER; PADÓ, 2018). Na WordNet, o MFS corresponde ao primeiro sentido de uma palavra.

Atualmente, as abordagens com melhor desempenho para o WSI se baseiam em substitutos lexicais (ALAGIĆ; ŠNAJDER; PADÓ, 2018; JURGENS; KLAPAFTIS, 2013). Nesta abordagem, um sistema cria um vetor substituto para cada palavra de destino a partir dos substitutos mais prováveis sugeridos por um modelo de linguagem estatística (NAVIGLI, 2012). Na Tarefa 13 do SemEval-2013 (JURGENS; KLAPAFTIS, 2013), o melhor desempenho dos sistemas WSI na configuração de sentido único é *F-score* = $0,64$. No (ALAGIĆ; ŠNAJDER; PADÓ, 2018), o desempenho do sistema usando o conjunto de dados SemEval-2010 é emparelhado *F-score* = 0.58. Na Tarefa 13 do SemEval-2015 (MORO; NAVIGLI, 2015), o melhor sistema para inglês é capaz de obter um desempenho semelhante aos sistemas mencionados acima.

3.3 MINERAÇÃO DE PADRÕES LINGUÍSTICOS E APLICAÇÕES

A mineração de padrões linguísticos em dados textuais é um processo que consiste em descobrir e extrair conhecimento de textos por meio de técnicas que identificam padrões e relacionamentos entre léxicos de uma língua (FELDMAN; SANGER *et al.*, 2007). Estes padrões podem apresentar diferentes formas e aplicações variadas. Em Silva (2017), padrões tal como: $< id : ID > < name : String > < type : StakeholderType >$ em que *ID* representa um identificador, *String* é um palavra (substantivo) e *StakeholderType* é uma pessoa, equipe ou sistema, são utilizados em linguagem de especificação de requisitos para desenvolvimento de sistemas.

Mendes e Antunes (2009) propõem outra representação para padrões linguísticos usando símbolos gramaticais entre colchetes, como o formato $[NP\ PRON]$, com o símbolo NP (frase nominal) seguido de $PRON$ (pronome). Tais padrões são utilizados para uma melhor compreensão da língua natural, o português neste caso. Este formato de padrão é minerado após o texto de entrada passar por etapas de processamento, tais como: análise morfológica, desambiguação morfológica, segmentação dos termos, NER, técnicas de decomposição e agrupamento (e.g. *chunking*).

Béchet *et al.* (2012) exploram padrões genéricos com correspondência (*matching*) a padrões específicos para detectar frases substantivas (*appositive qualifying phrases*), em língua francesa. Por exemplo, o padrão genérico

$\langle(NOUN)(PRP)(NOUN)\rangle$, que representa léxicos classificados em sequência com os símbolos *NOUM* (substantivo) e *PRP* (preposição), combinam com os padrões específicos $\langle(homemNOUN)(dePRP)(NOUN)\rangle$ e $\langle(homemNOUN)(dePRP)(convicoNOUN)\rangle$.

Quiniou *et al.* (2012) também exploram padrões linguísticos usando classificação gramatical, como em Béchet *et al.* (2012), porém apresentam padrões de itens simples para a análise estilística de texto. Neste caso, os padrões linguísticos são constituídos pelas palavras e o símbolo *, que representa um espaço de uma ou mais palavras. Por exemplo, o padrão de item simples *some*more*than* tem como uma instância correspondente a frase "*he has **some** bites **more** venomous **than** those from your mouth*".

Outra linha de pesquisa relevante é WSI/WSD baseada na detecção de padrões lexicais. As abordagens baseadas em padrão lexical para WSD exploram dependências gramaticais (por exemplo, sintática, morfológica, semântica) de palavras e contextos, ou pares e padrões de palavras. Por exemplo, o uso de estruturas de dependência sintática derivadas automaticamente do texto para criar uma coleção de caminhos que compartilham âncoras nominais semelhantes à distribuição (LIN; PANTEL, 2001a); aprender uma combinação ideal das várias fontes de conhecimento para palavras-chave individuais (MIHALCEA, R. F., 2002); abstração de padrões usando recursos (por exemplo, lema e parte do discurso) e métodos de kernel (STRAPPARAVA; GLIOZZO; GIULIANO, 2004); identificar padrões lexicais entre conceitos em uma enciclopédia *on-line* para estender ontologias existentes ou redes semânticas (RUIZ-CASADO; ALFONSECA; CASTELLS, 2007); usando conjuntos de classificadores Bayesianos ingênuos para extrair instâncias sensoriais infrequentes com o mesmo n-grama (LIU, P.-Y. *et al.*, 2010). Todos esses trabalhos apresentam uma avaliação dos resultados com medidas clássicas de precisão, cobertura e medida-F usando dados do SemEval (exceto em (LIN; PANTEL, 2001a)).

A disponibilidade dos conjuntos de dados SemEval aprimorou os trabalhos no WSI/WSD com base em padrões lexicais (MIHALCEA, R. F., 2002; STRAPPARAVA; GLIOZZO; GIULIANO, 2004; LIU, P.-Y. *et al.*, 2010). Por exemplo, a Tarefa 13 do SemEval-2013[2] forneceu três conjuntos de dados com anotações de sentido de palavra ponderadas e rotuladas com sentidos do WordNet 3.0. Apesar da cobertura dos conjuntos de dados SemEval em diferentes domínios e idiomas, a maioria dos conjuntos de dados relacionados à tarefa WSD são textos de documentos, poucos deles estão no domínio de mídia social[3] (por exemplo, expressões com várias palavras e supersensores de substantivos e verbos) (SCHNEIDER *et al.*, 2016).

Sorato e Fileto (2019) exploram padrões de linguagem que são análogos aos explorados neste trabalho, mas para análise de discurso e classificação de texto. Os

[2] `https://www.cs.york.ac.uk/semeval-2013/task13/`, textos extraídos de tipos de fontes como ficção, diário, cartas, guias de viagem técnicos e não-ficção.

[3] `http://dimsum16.github.io/`

padrões SSP *Short Semantic Pattern (SSP)* referem-se a sequências de palavras que são semanticamente semelhantes entre si e recorrentes em microblogs, como *tweets*. O método proposto recebe como entrada os textos a serem analisados e um conjunto de palavras que indicam os alvos dos discursos de interesse. Eles empregam combinações de palavras previamente treinadas com técnicas de aprendizado profundo para identificar partes do texto centralizadas nas palavras fornecidas (ou similares) e que correspondem (em termos de seus significados, e não suas estruturas lexicais ou sintáticas) a outras sequências de texto. Em outras palavras, um SSP pode ser considerado como um componente semântico exclusivo que é expandido de uma palavra de destino para abranger os vizinhos, mantendo a similaridade semântica (calculada usando vetores de palavras) com outros componentes expandidos. A aplicação do método em *tweets* enviados por "*Donald Trump*" para falar sobre seus oponentes permitiu revelar, entre outros resultados em diferentes conjuntos de dados, a identificação de semânticas recorrentes em seus discursos. Por exemplo, muitos dos padrões de texto extraídos mostram que *Trump* frequentemente retrata seus oponentes na mídia como mentirosos, corruptos e incompetentes. As representações baseadas no SSP do conteúdo do texto foram fornecidas como características aos classificadores de discursos, levando a ganhos leves, mas consistentes, no desempenho da classificação do texto.

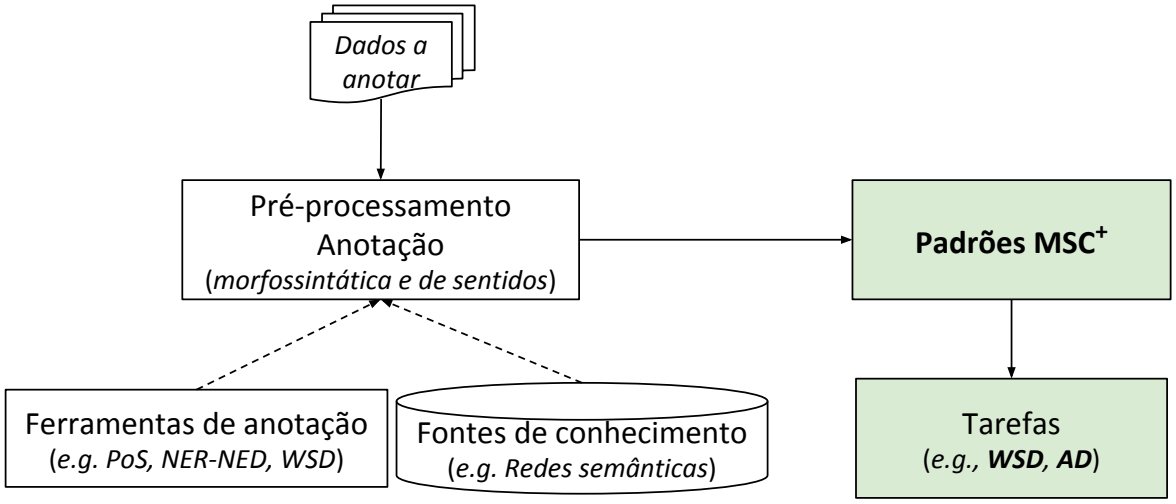
A abordagem proposta nesta tese combina várias das ideias acima e adiciona recursos que garantem a interpretabilidade, pois são usados inventários com oos sentido de palavras (ALAGIĆ; ŠNAJDER; PADÓ, 2018) e uma ferramenta tagger/WSD para enriquecer os dados das mídias sociais com informações semânticas; para induzir o sentido, contamos com recursos de contexto morfossintático e semântico (STRAPPARAVA; GLIOZZO; GIULIANO, 2004; PANCHENKO *et al.*, 2017), coocorrências (PANCHENKO *et al.*, 2017) e aprendizado de padrões (MIHALCEA, R. F., 2002; RUIZ-CASADO; ALFONSECA; CASTELLS, 2007; LIU, P.-Y. *et al.*, 2010). O método WSI/WSD proposto também pode ser adaptado para uso em domínios de aplicativos específicos, usando diferentes conjuntos de dados e bases de conhecimento, como vocabulários médicos (MONDAL *et al.*, 2015). Além disso, é apresentado um experimento complementar usando a abordagem proposta inicialmente para WSI/WSD, na análise do discurso/conteúdo de textos literários em português.

4 ABORDAGEM MSC^+

Neste capítulo são apresentadas as definições formais (Seção 4.1) para induzir e/ou desambiguar o sentido correto de palavras em texto por meio de padrões linguísticos. A abordagem proposta parte da hipótese de que tarefas como indução ou desambiguação lexical do sentido de palavras e análise do discurso/conteúdo podem se beneficiar de informações contextuais fornecidas por sequências de Componentes Morfo-Semânticos (MSC).

A Figura 10 ilustra as principais etapas que compõem a abordagem MSC^+. A primeira etapa da abordagem proposta recebe como entrada um conjunto de dados a anotar que passa por tarefas de pré-processamento e anotação morfossintática e de sentidos, sendo dependente de diferentes fontes de conhecimento e ferramentas de anotação.

Figura 10 – Arquitetura da abordagem proposta.



Na segunda etapa, as informações de contexto de alguma coleção de texto curto são captadas automaticamente por um algoritmo não supervisionado (Seção 4.1.2) que extrai padrões da linguagem definidos por sequências de MSC, chamados de padrões MSC^+. Cada MSC possui os sentidos mais prováveis (por exemplo, conceitos candidatos ou entidades nomeadas) e as classes morfossintáticas mais prováveis (e.g. marcações de *PoS)*) para cada palavra relevante em um documento de texto. Um padrão MSC^+ pode ser visto como uma sequência de pares $\langle$*classe morfossintática, sentido*$\rangle$ que aparecem repetidamente nos MSCs de textos curtos distintos (e.g. *tweets*). Uma instância de um padrão MSC^+ é uma sequência de MSCs consecutivos, e não necessariamente adjacentes. Em síntese, uma sequência de MSCs pode ser entendida como palavras consecutivas em texto, em que as classes morfossintáticas e os sentidos mais provável para cada MSC correspondem àquelas na respectiva posição

do padrão. Na etapa final da abordagem proposta, os padrões MSC^+ são utilizados para uma tarefa específica, tais como a desambiguação do sentido de palavras (WSD), análise/classificação do discurso (AD).

4.1 FORMALIZAÇÃO

Algumas palavras podem ter seus significados induzidos e desambiguados por meio de padrões morfo-semânticos. Um padrão morfo-semântico é uma sequência de Componentes Morfo-semânticos (*Morpho-Semantic Components - MSC*) com o tamanho variando em n-grama. Um MSC refere-se a uma palavra de um documento d que possui relação com um conjunto de candidatos: classes morfossintáticas e de sentidos, como segue na Definição 1.

Definição 1 . ***Componente Morfo-semântico - MSC***. *Dado um documento d com identificador idd, um conjunto de classes morfossintáticas Λ e um conjunto de sentidos Ξ, um MSC é uma sêxtupla $msc = \langle idd, idc, b, e, P, S\rangle$, onde:*

$idd:$ *é o identificador do documento (e.g., id de um tweet) onde o msc ocorre;*

$idc:$ *é o identificador do componente morfo-semântico msc;*

$b:$ *é o início do msc identificado no documento idd;*

$e:$ *é o final do msc identificado no documento idd;*

$P = \{\langle p_1, \tilde{p}_1\rangle, \ldots, \langle p_n, \tilde{p}_n\rangle\}$ *($n \geq 1$): é uma lista de n classes de PoS candidatas para o msc, em que cada par $\langle p_i, \tilde{p}_i\rangle$ indica uma classe candidata $p_i \in \Lambda$ com seus graus de crenças normalizados $\tilde{p}_i \in (0, 1]$, sob a forma de $\sum_{i=1}^{n} \tilde{p}_i = 1\ \forall \langle p_i, \tilde{p}_i\rangle, \langle p_j, \tilde{p}_j\rangle \in P : p_i \neq p_j$;*

$S = \{\langle s_1, \tilde{s}_1\rangle, \ldots, \langle s_m, \tilde{s}_m\rangle\}$ *($m \geq 1$): é uma lista de m classes de sentidos candidatos para msc, em que cada par $\langle s_i, \tilde{s}_j\rangle$ indica um sentido candidato $s_j \in \Xi$ com seu grau de crença normalizado $\tilde{s}_j \in (0, 1]$, sob a forma de $\sum_{j=1}^{m} \tilde{s}_j = 1$ $\forall \langle s_i, \tilde{s}_i\rangle, \langle s_j, \tilde{s}_j\rangle \in S : s_i \neq s_j$.*

Para cada candidato, classe morfossintática $p_i \in \Lambda$ e sentido s_j é associado um grau de confiança ($\tilde{p}_i$ e $\tilde{s}_j$, respetivamente), porque pode haver incertezas sobre as classes de *PoS* e os sentidos (semânticas possíveis) de cada MSC retornado nas anotações realizadas por ferramentas. O conjunto de possíveis classes de *PoS* e sentidos de palavras variam de acordo com a ferramenta de anotação e o inventário de sentidos utilizados na anotação do texto. Por exemplo, classes de ontologia, recursos de dados ligados e *synsets* de léxicos são comumente usados para representar os sentidos das palavras. Quando o valor anotado de uma palavra é uma instância, o

sentido pode ser induzido ou comparado com outros sentidos referentes à sua classe (por exemplo, marca de cerveja, comida, pessoa, instituição).

Certamente, existem correlações entre as classes morfossintáticas e os sentidos das palavras, pois um certo sentido só se aplica a algumas palavras quando elas desempenham papéis morfossintáticos específicos nas frases. No entanto, padrões linguísticos sob a forma de MSCs separam os sentidos e classes morfossintáticas em virtude de dois motivos: **(i)** existem ferramentas que anotam as palavras apenas com os sentidos e classes morfossintáticas que são conhecidos (ou mais frequentes); **(ii)** o dinamismo da linguagem pode dificultar o controle de todos os sentidos e classes morfossintáticas que podem ser associadas a certas palavras e as correlações entre as respectivas classes.

Um MSC de um texto d é considerado relevante se w é considerada relevante para expressar o conteúdo de d segundo algum método de sumarização aplicado a d, por exemplo, entre outras possibilidades. Os MSCs também podem ser classificados em não-desambiguado, fracamente desambiguado ou fortemente desambiguado, dependendo da confiança de seus sentidos e classes morfossintáticas candidatos. Definição 2 e 3 declara as condições para um MSC ser considerado fracamente desambiguado e fortemente desambiguado, respectivamente.

Definição 2 . ***MSC fracamente desambiguado*** *Um componente morfo-semântico msc é fracamente desambiguado em relação aos parâmetros τ^p_{abs}, $\tau^s_{abs} \in (0,1]$ se, e somente se, existirem pelo menos um par $\langle p, \tilde{p} \rangle \in P$ e ao menos um par $\langle s, \tilde{s} \rangle \in S$ de tal modo que:*

$$\tilde{p} \geq \tau^p_{abs} \quad \wedge \quad \tilde{s} \geq \tau^s_{abs}$$

Definição 3 . ***MSC fortemente Desambiguado*** *Um componente morfo-semântico msc é fortemente desambiguado em relação aos parâmetros τ^p_{abs}, τ^s_{abs}, τ^p_{dif}, $\tau^s_{dif} \in (0,1]$ se, e somente se, está fracamente desambiguado em relação à τ^p_{abs} e τ^s_{abs}, e existir um par $\langle p, \tilde{p} \rangle \in P$ e um par $\langle s, \tilde{s} \rangle \in S$ tal que:*

$$\forall \langle p_i, \tilde{p}_i \rangle \in P | p_i \neq p : \tilde{p} - \tilde{p}_i \geq \tau^p_{dif} \quad \wedge$$
$$\forall \langle s_j, \tilde{s}_j \rangle \in S | s_j \neq s : \tilde{s} - \tilde{s}_j \geq \tau^s_{dif}$$

Um MSC fortemente desambiguado tem apenas um sentido candidato no qual a confiança supera a de todos os outros candidatos em pelo menos τ^s_{dif} e uma classe morfossintática na qual a confiança supera as de todas as outras classes morfossintáticas em pelo menos τ^p_{dif}. A representação de um MSC fortemente desambiguado pode ser simplificada para incluir apenas a classe morfossintática desambiguada p e o sentido desambiguado s, da seguinte forma: $msc_desambiguado = \langle idd, idc, b, e, p, s \rangle$.

4.1.1 Padrão MSC^+

Muitos MSCs relevantes podem aparecer em um documento de texto d, possivelmente separados um do outro por palavras consideradas irrelevantes (e.g. *stopwords*). A definição 4 especifica formalmente uma sequência de MSCs em um documento de texto.

Definição 4 . ***Sequência*** MSC^+. *Uma sequência de componentes morfo-semânticos (sequência $MSC+$) é uma lista ordenada $msc^+ = (msc_1, \ldots, msc_l)$ ($l \geq 1$) de componentes morfo-semânticos referentes a palavras subsequentes, mas não necessariamente adjacentes, no mesmo documento d, i.e., $\forall msc_k = \langle idd, idc_k, b_k, e_k, P_k, S_k \rangle, msc_{k+1} = \langle idd, idc_{k+1}, b_{k+1}, e_{k+1}, P_{k+1}, S_{k+1} \rangle \in msc^+ : b_{k+1} > e_k \ (k = 1, \ldots, l)$.*

Uma sequência MSC^+ é considerada fortemente desambiguada se todos os seus componentes morfo-semânticos $l \geq 1$ estiverem fortemente desambiguados de acordo com o conjunto de parâmetros τ^p_{abs}, τ^s_{abs}, $2\tau^p_{dif}$, $\tau^s_{dif} \in (0, 1]$. As sequências MSC^+ com o mesmo comprimento (l) em que os seus respectivos MSCs são todos fortemente desambiguados para a mesma classe de *PoS* e sentido, constituem ocorrências (instâncias) de um mesmo padrão MSC^+, como especificado nas Definições 5 e 7.

Definição 5 . ***Padrão*** MSC^+. *Um padrãoMSC^+ é uma lista ordenada de $l \geq 1$ pares $pmsc^+ = (\langle p_1, s_1 \rangle, \ldots, \langle p_l, s_l \rangle)$, onde cada par $\langle p_k, s_k \rangle$ ($1 \leq k \leq l$) refere-se a uma classe morfológica ou morfossintática p_k e a um sentido s_k.*

Definição 6 . ***Instância de correspondência perfeita de um padrão*** MSC^+. *Considerando um padrão MSC^+ $pmsc^+ = (\langle p_1, s_1 \rangle, \ldots, \langle p_l, s_l \rangle)$ de comprimento $l \geq 1$, uma instância de correspondência perfeita com um $pmsc$ é qualquer sequência MSC^+, $msc^+ = (msc_1, \ldots, msc_l)$ também de comprimento l, na qual cada componente morfo-semântico $msc_k = \langle idd_k, idc_k, b_k, e_k, p_k, s_k \rangle$ ($1 \leq k \leq l$) de msc^+ está fortemente desambiguado, conforme os parâmetros τ^p_{abs}, τ^s_{abs}, τ^p_{dif}, $\tau^s_{dif} \in (0, 1]$, à respectiva classe morfossintática p_k e sentido s_k (pares $\langle p_k, s_k \rangle$) de $pmsc^+$.*

As instâncias distintas de um padrão MSC^+ podem aparecer em documentos distintos. Fato este apresentado nos *tweets* da Tabela 5 (Seção 2.5.2), cada um com uma instância do padrão $\langle VER, Ingerir \rangle, \langle SUB, Cerveja \rangle, \langle ADJ, Qualquer \rangle$.

Finalmente, uma sequência MSC MSC^+ de um documento de texto d corresponde parcialmente a um padrão MSC $pmsc^+$ quando pelo menos um componente msc_k da sequência MSC^+ corresponde ao sentido e a classe morfossintática do respectivo par $langlep_k, s_k\ rangle$ do padrão $pmsc^+$, conforme declarado em Definição 7.

Os sentidos dos componentes MSC restantes de msc^+ estão indefinidos (sem sentidos candidatos) ou pelo menos não fortemente desambiguados.

Definição 7 . ***Instância de correspondência parcial de um padrão*** MSC^+. *Considerando um padrão MSC^+ $pmsc^+ = (\langle p_1, s_1\rangle, \ldots, \langle p_l, s_l\rangle)$ de comprimento $l \geq 1$ e uma sequência MSC^+ $msc^+ = (msc_1, \ldots, msc_l)$ também de comprimento l l, uma correspondência parcial de um padrão MSC^+ é qualquer sequência $pmsc^+$ se ao menos um componente MSC de MSC^+ está fortemente desambiguado, de acordo com os parâmetros τ^p_{abs}, τ^s_{abs}, τ^p_{dif}, $\tau^s_{dif} \in (0, 1]$, à respectiva classe morfossintática e sentido de $pmsc^+$, i.e., $\langle p_k, s_k\rangle$, e ao menos uma classe morfossintática $p_{k,i} \in P_k$ ($1 \leq k \leq l, i \geq 1$) de cada MSC em MSC^+ corresponde a um respectivo par $\langle p_k, s_k\rangle$ do padrão $pmsc$.*

4.1.2 Mineração de padrões MSC^+

O algoritmo 1 descreve o método proposto para minerar padrões MSC^+ e associar cada um deles às suas respectivas instâncias de correspondência parcial ou perfeita. O algoritmo assume como entradas os componentes morfo-semânticos (MSCs) extraídos de um conjunto de documentos de texto e um valor de suporte mínimo $sup \geq 1$ (número de instâncias) usado para retornar os padrões minerados.

Algorithm 1: Mining MSC^+ patterns

```
   input  : mscData; // lista dos MSCs encontrados em documentos
   input  : sup; // suporte mínimo
   output : Padrões MSC+ com suas instâncias.
1  begin
2      foreach p ∈ P do
3          foreach s ∈ S do
4              matchings ← filter msc ∈ mscData matching ⟨p, s⟩;
5              if matchings.size ≥ sup then
6                  matchingMSCs[p, s] ←matchings;
7      mscMines(mscData, matchingMSCs, sup);
8  mscMines (mscData, matchingMSCs, sup):
9      foreach k, v ∈ mscData do
10         foreach i, c ∈ matchingMSCs do
11             if k ≠ i then
12                 mscIntersection ← v ∩ c;
13                 if mscIntersection ≠ ∅ then
14                     mscMine[k, i] ← mscIntersection;
15     if mscMine ≠ ∅ then
16         fileOut(mscMine);
17         mscMines(msc, mscMine, sup, n);
```

A primeira chamada do algoritmo leva todos os MSCs com sentidos resolvidos a partir de um conjunto de dados de treinamento. Posteriormente, o algoritmo pode ser chamado para encontrar correspondências para MSCs (geralmente de novos documentos) cujo sentido deve ser induzido e/ou desambiguado. Os MSCs das sentenças são extraídos automaticamente e rotulados no formato de dados vertical (FOURNIER-VIGER *et al.*, 2017). Apenas os MSCs com classes morfossintática e sentidos compatíveis são considerados para a correspondência parcial ou perfeita de um padrão. Uma correspondência parcial ocorre quando, em algum documento, um ou mais MSCs que correspondem a um padrão, possui um dos seus MSCs com o sentido não definido ou ambíguo (fracamente desambiguado).

Ainda em relação ao algoritmo 1, nas linhas 1 a 6 são analisadas as incidências de combinações entre as marcações morfossintáticas e sentidos. Os padrões com o comprimento um (no formato $\{\langle p, s\rangle\}$) são gerados a partir de remoções com base no suporte sup. As instâncias (com um MSC cada) são armazenadas no vetor $matchingMSCs$. Então, os padrões são extraídos (função $mscMines$) de comprimento maior que um de maneira recursiva, comparando o conteúdo dos IDs de candidatos *msc* ($v \cap c$, com vetor de alinhamento à esquerda). Assim, uma sequência de componentes MSC^+ tem alta relevância quando é suportada por várias instâncias em que a classe morfossintática e sentido são compatíveis com os de um par $\langle p, s\rangle$. Essa função também implementa a remoção com base em valores de suporte, filtros para selecionar conteúdo e verificar a ordem dos componentes.

Os padrões MSC^+ e às instâncias resultantes do algoritmo 1 são retornados em um arquivo no formato JSON. A Tabela 5 (Seção 2.5.2) apresenta duas instâncias de correspondência perfeitamente ao padrão de tamanho três: $\{\langle VER, Ingerir\rangle, \langle SUB, Cerveja\rangle, \langle ADJ, Qualquer\rangle\}$ ocorrendo nos *tweets* com os identificadores #1 e #2, respectivamente. Esse padrão pode ser usado para induzir e/ou desambiguar o sentido de palavras. Como as outras instâncias que correspondem parcialmente a esse padrão (*tweets* identificados entre #3 e #5) podem ter o segundo componente sendo desambiguado para Cerveja, o suporte ao padrão aumenta.

4.1.3 WSD por meio de padrões MSC^+

O algoritmo 1 (Seção 4.1.2) associa dois tipos de instâncias (sequências MSC encontradas nos documentos de texto) aos respectivos padrões MSC^+: (**i**) **instâncias de correspondência perfeita**, que ocorre quando todos os MSCs de uma sequência estão fortemente desambiguados de acordo com a respectiva classe morfossintática e sentido $\langle p, s\rangle$ de algum padrão MSC^+, conforme definido na Seção 4.1; e (**ii**) **instâncias de correspondência parcial**, que são sequências de MSCs nas quais pelo menos um dos componentes possui o seu sentido indefinido, não desambiguado, ou ainda fracamente desambiguado. Assim, se outros componentes de uma instância

de correspondência parcial aderirem perfeitamente ao par $\langle p, s\rangle$ do padrão MSC^+, o(s) sentido(s) indefinido(s) podem ser induzidos e desambiguados de acordo com o sentido do respectivo par $\langle p, s\rangle$ do padrão MSC^+.

Por exemplo, os sentidos das palavras em MSC_2 nos *tweets* #3 (**loira**), #4 (**devassa**) e #5 (**polar**) da Tabela 5 (Seção 2.5.2), podem ser desambiguados para **Cerveja**, graças à correspondência parcial dos respectivos MSCs em sequências com o padrão $\{\langle VER, Ingerir\rangle, \langle SUB, Cerveja\rangle, \langle ADJ, Qualquer\rangle\}$. Dessa forma, quanto maior o suporte (número de instâncias MSC com correspondência perfeita ao padrão), menor o número de componentes fracamente desambiguados nas sequências MSC de correspondência parcial, e maior a confiança para induzir e desambiguar o sentido.

Quando uma palavra com o sentido indefinido aderir as sequências MSC^+ que correspondem parcialmente a mais de um padrão, métodos com base em escore podem ser utilizados para inferir o significado de palavras. Tais palavras podem ser combinadas com os pares classe morfossintática e sentido ($\langle p, s\rangle$) na posição do respectivo MSC em cada padrão, considerando os valores de crença $\tilde{p}$ e $\tilde{s}$ dos respectivos componentes MSCs nas instâncias de correspondência perfeita ao padrão. Os métodos para a indução e desambiguação do sentido propostos nesta tese são:

1. **Máxima adesão média ao sentido (MAA-S)**: este método induz o sentido de uma palavra alvo w_t em uma sequência MSC^+ de correspondência parcial a mais de um padrão MSC^+, ao qual aderem instâncias fortemente desambiguadas ao padrão MSC^+. O método MAA-S calcula a pontuação de cada padrão ($Score$) e, consequentemente, o sentido s do MSC correspondente, considerando a média das crenças $\tilde{s}$ do respectivo MSC fortemente desambiguado de cada instância de correspondência perfeita ao padrão i, conforme define a Equação 2.

2. **Máxima aderência média a classe morfossintática e sentido (MAA-PS)**: O método MAA-PS calcula um escore de indução usando as crenças $\tilde{p}$ e $\tilde{s}$ de cada classe morfossintática p e sentido s candidatos do respectivo MSC fortemente desambiguado (correspondendo à posição da palavra w_t) em cada instância que corresponde perfeitamente ao padrão. A pontuação de cada w_t para cada padrão i é calculada de acordo com a Equação 3.

$$Score(w_{ti}) = \frac{\sum_{j=1}^{sup_i} \tilde{s}_{ji}}{sup_i} \tag{2}$$

$$Score(w_{ti}) = \frac{\sum_{j=1}^{sup_i} (\tilde{p}_{ji} + \tilde{s}_{ji})}{2\ sup_i} \tag{3}$$

Nas Equações 2 e 3, sup_i é o número total de sequências MSC^+ que correspondem perfeitamente a um padrão MSC^+ i. Então, o sentido para w_t é determinado

a partir da posição correspondente do padrão com a pontuação máxima entre os candidatos ($BestFit$), conforme declarado na Equação 4.

$$BestFit = \underset{i}{\operatorname{argmax}}\, Score(w_{ti}) \tag{4}$$

Conforme apresentado neste capítulo, abordagens para o aprendizado de padrões linguísticos em texto curto são alternativas para explorar o significado de palavras em aplicações específicas, como as baseadas no processamento da linguagem existente em mídias sociais e na web de uma maneira geral. No próximo capítulo são descritos os experimentos realizados para verificar a efetividade da abordagem proposta nas tarefas de WSD e na Análise Automática do Discurso. A performance da abordagem proposta é avaliada em diferentes conjuntos de texto e idiomas, considerando os diferentes níveis de complexidade semântica nas tarefas selecionadas. Também é investigado o impacto causado pelo uso de fontes externas de conhecimento, tanto para o enriquecimento semântico dos conjuntos de textos quanto para a melhoria da performance do algoritmo não supervisionado de mineração de padrões MSC^+.

5 USO DE PADRÕES MSC^+ EM WSD

Algumas abordagens para induzir e/ou desambiguar o sentido de palavras utilizam o aprendizado de padrões linguístico (LIU, P.-Y. *et al.*, 2010). No entanto, os padrões MSC^+ podem ser considerados como uma alternativa para algumas situações em que as abordagens atuais falham, como exemplificado no capítulo anterior (Seção 2.5.2). Os padrões linguísticos, como o proposto nesta tese, podem ser utilizados para aprimorar as anotações semânticas produzidas por diferentes métodos de anotação (MORO; RAGANATO; NAVIGLI, 2014; FILETO *et al.*, 2015), e consequentemente impulsionar uma variedade de tarefas e aplicativos computacionais, tais como simplificadores de texto (SAGGION; ŠTAJNER *et al.*, 2015), sumarizadores (GOULARTE; NASSAR *et al.*, 2019), enriquecimento de base de conhecimento (FELLBAUM, 1998; CAMACHO-COLLADOS; PILEHVAR; NAVIGLI, 2016; NAVIGLI; PONZETTO, 2010; A. JÚNIOR; SCHIEL; MARINHO, 2015), análise de sentimentos (DRAGONI, 2018) e resposta a perguntas (AL-HARBI; JUSOH; NORWAWI, 2017), etc.

Este capítulo apresenta o experimento realizado com a abordagem MSC^+ para capturar o sentido de palavras com semântica implícita. Inicialmente, ele fornece uma visão geral do fluxo de informações que compõem os processos selecionados para explorar a abordagem proposta na tarefa de WSD, para depois detalhar a descrição de como tais processos podem ser realizados. Além da organização e da configuração dos experimentos, que descrevem as bases de dados utilizadas, também são apresentados e discutidos os impactos causados no desempenho do algoritmo de mineração de padrões MSC^+ por fatores como o tipo de escrita e idioma.

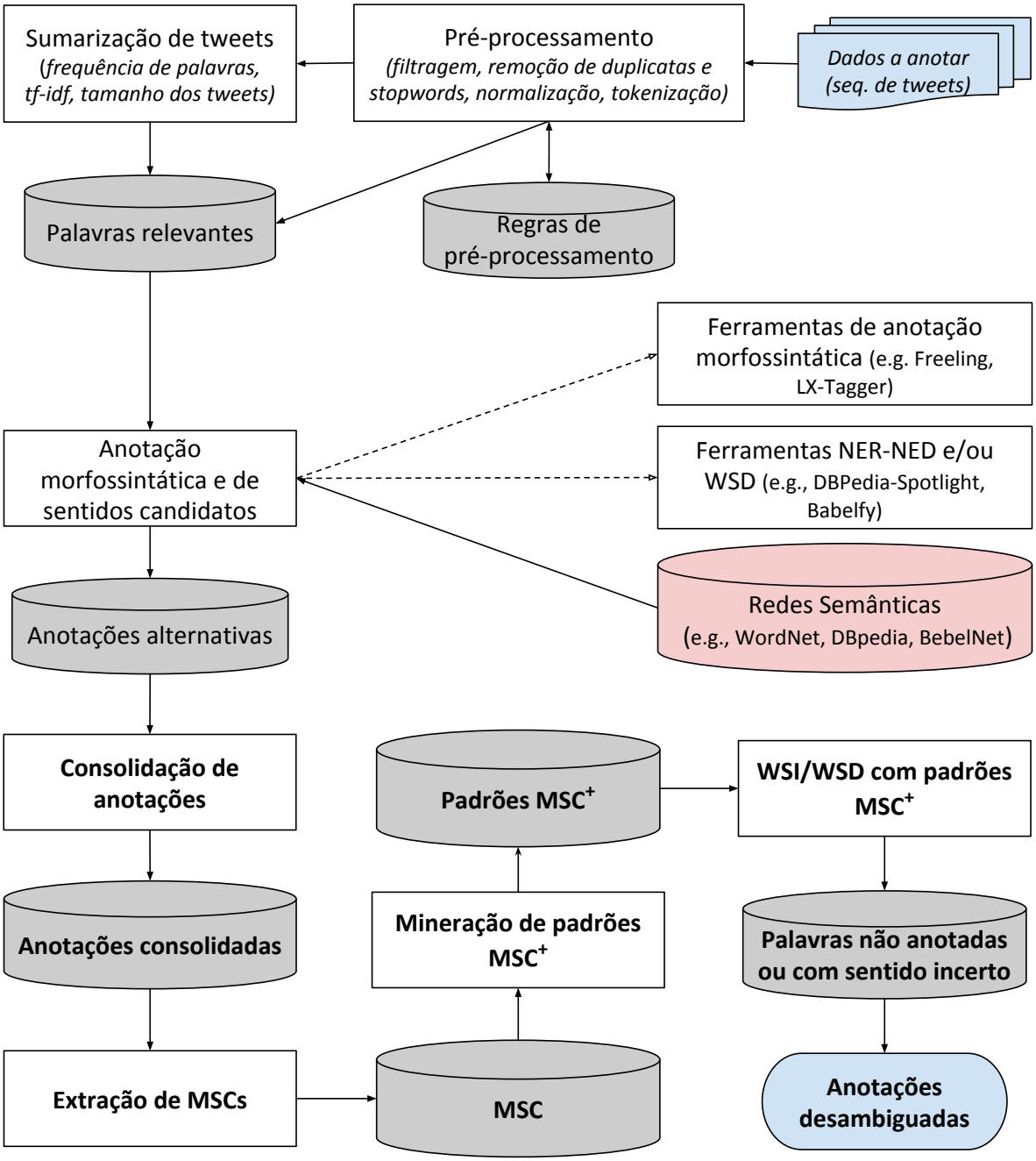
A Seção 5.1 descreve a infra-estrutura do experimento, incluindo ferramentas utilizadas, os parâmetros de funções e métricas de avaliação. O primeiro experimento tem como objetivo investigar a prevalência de padrões MSC^+ em *tweets*, seu suporte combinando perfeitamente as sequências MSC^+ nesses *tweets* e os ganhos obtidos na tarefa de WSI/WSD com base nos padrões minerados em palavras com sentido não definido pelas ferramentas de anotação típicas.

5.1 DESCRIÇÃO DO EXPERIMENTO

Este experimento descreve a abordagem MSC^+ para manipular instâncias desafiadoras dos problemas de indução e desambiguação lexical de sentido em sequências de textos, como postagens em mídias sociais. A Figura 11 fornece uma visão geral do processo proposto para o experimento, composto por sete etapas: (i) `Pré-processamento`, (ii) `Sumarização de` *`tweets`* (iii) `Anotação morfossintática e de sentidos candidatos`, (iv) `Consolidação de anotações`, (v) `Extração de MSCs`, (vi) `Mineração de padrões` MSC^+, (vii) `WSI/WSD com padrões` MSC^+. Na figura, as linhas contínuas indicam o fluxo de dados – o resultado de cada etapa que é

armazenado, e as linhas pontilhadas vinculadas na etapa (iii) com as ferramentas de classificação morfossintática (e.g., FreeLing, LX-Tagger) e WSI/WSD e/ou NER/NED (e.g., DBpedia-Spotlight, Babelfy) indicam a função dependência do uso de tais ferramentas, portanto a etapa (iii) é dependente dessas subetapas. As subetapas mencionadas podem ser acessadas remotamente por servidores da Web por meio de APIs. Muitas destas ferramentas contam com inventários do sentido do tipo léxico (e.g. WordNet), coleções de dados ligados (LOD) (e.g. DBpedia) ou composições deles (e.g. Babelnet), geralmente representados como grafos de conhecimento.

Figura 11 – Abordagem proposta para WSD.



5.1.1 Base de dados utilizada

Para permitir a análise empírica da abordagem MSC^+ e testar diferentes configurações para os métodos WSI/WSD propostos, foi selecionado o conjunto de dados Microposts2016[1]. O Microposts2016 consiste em *tweets* no idioma inglês de eventos e não-eventos extraídos de uma coleção de mais de 18 milhões de *tweets*. Após a tarefa de filtragem (remoção dos tweets "*Not Available*" – àqueles que não mais estão no servidor do Twitter), foram obtidos 2493 tweets, com 827 menções anotadas a entidades no padrão ouro. A Tabela 7 apresenta alguns exemplos de postagens deste conjunto de dados.

Tabela 7 – Uma amostra de *tweets* do conjunto de dados.

Exemplo de posts:	
676782075995787264	RT @user: Star Wars theforceawakens first reactions https://t.co/8KZno5CW49
676802348409229313	Fans who saw the world premiere of StarWars TheForceAwakens say it did not disappoint https://t.co/J7pCNS5xt6
676785373943611393	my body is ready TheForceAwakens https://t.co/ThEcn9p6OF
Exemplo de entidade anotada:	
674962184615682048	55 64 http://dbpedia.org/resource/Star_Wars 1 Product

5.1.2 Pré-processamento e sumarização de *tweets*

O pré-processamento dos dados de entrada tem o objetivo de minimizar problemas encontrados em textos de mídias sociais, como os detalhados na Seção 2.6. As postagens em mídia social são muito propensas a ruído, sendo comum apresentarem *emoticons/emoji*, conteúdo em diferentes idiomas, o uso incorreto de letras maiúsculas e minúsculas, falta de vogais/consoantes em abreviação de palavras e acrônimos, gírias, caracteres não reconhecidos por *parsers* das ferramentas de anotação (por exemplo, @ = *usuário*, # = *hashtag*, *url*) e outros problemas. Portanto, antes de processar o texto, é necessário submeter as mensagens à tarefas de limpeza, filtragem e normalização de seus textos. A limpeza remove as partes menos úteis do texto, como palavras irrelevantes (*stopwords*). A filtragem, por outro lado, pode selecionar as partes mais relevantes, com base em palavras-chave, tópicos ou idioma (e.g., inglês, português, espanhol). A normalização pode trocar palavras fora do vocabulário com palavras formais equivalentes de um repositório.

[1] `http://microposts2016.seas.upenn.edu/challenge.html`

Nesta tese, foram consideradas como partes úteis do texto apenas símbolos que adicionam informações relevantes à mensagem, tais como verbos, substantivos, advérbios e adjetivos. A Tabela 8 ilustra um exemplo de texto bruto de *tweet* (entrada) e como ele é transformado após o estágio de pré-processamento (saída).

Tabela 8 – Um exemplo de pré-processamento de *tweet*.

Entrada:	RT @user: Star Wars #theforceawakens first reactions https://t.co/8KZno5CW49.
Saída:	*user Star Wars theforceawakens reactions*

Os textos publicados em mídias sociais, geralmente possuem poucos dados de contexto. No entanto, pode-se adaptar técnicas de sumarização de texto a *tweets* tendo como finalidade enriquecer o contexto das postagens. As métricas de sumarização de texto, por exemplo, podem ser aplicadas para selecionar uma coleção de postagens de certos *usuários*, considerando apenas *hashtags* de interesse, ou aquelas originárias de uma determinada região e/ou cujo *timestamp* esteja em um determinado intervalo de tempo. Os textos destas postagens tendem a apresentar características de linguagem semelhantes e, considerados como coleções, podem fornecer mais informações contextuais. A abordagem MSC^+ tira proveito deste fato para extrair padrões de linguagem que são comuns em coleções de postagens e usá-los para induzir e dasambiguar o sentido de palavras.

5.1.3 Anotação morfossintática e de sentidos candidatos

A anotação morfossintática e de sentidos candidatos é um processo que tem por objetivo enriquecer os elementos textuais resultantes do estágio de pré-processamento com informações semânticas (e.g., bases de conhecimento semântico) e marcações de PoS (Ferramentas de classificação morfossintática, NER/NED, WSD).

Para realizar esta etapa, selecionou-se duas ferramentas de anotação morfossintática e de sentidos com suporte aos idiomas inglês e português, a Freeling e Babelfy. A **FreeLing** fornece funcionalidades de análise de linguagem úteis para construir o MSC, como tokenização de texto, análise morfológica, detecção e classificação de entidades nomeadas, marcação de PoS, anotação de sentido baseada na WordNet. O FreeLing usa o UKB, um sistema WSD de ponta, baseado no PageRank personalizado (AGIRRE; SOROA, 2009). A **Babelfy** é uma ferramenta das principais ferramentas do estado-da-arte para WSD e vinculação de entidades em textos escritos em várias línguas. A Babelfy utiliza a BabelNet 4.0, uma grande rede semântica multilíngue que conecta descrições de conceitos e objetos de diferentes inventários, como WordNet,

Open Multilingual WordNet, Wikipedia, OmegaWiki, Wikcionário e Wikidata (NAVIGLI; PONZETTO, 2010).

A Tabela 9 apresenta anotações geradas pela ferramenta FreeLing[2] sobre componentes textuais relevantes resultantes do pré-processamento de texto ilustrado com o primeiro *tweet* da Tabela 8. Conforme a Tabela 9, a Freeling associa a cada símbolo (primeira coluna) seu lema (segunda coluna) e sua classe morfossintática (terceira coluna). No caso da classe morfossintática, `NP00V00 1` significa que a ferramenta classificou *Star_Wars* como um nome próprio cujo gênero e número não são especificados e cuja classe de entidade nomeada é *other*, com confiança 1 (100 %). O `VBZ` significa que *theforceawakens* foi classificado como verbo na terceira pessoa com confiança 0.9967. Finalmente, `NNS 1` significa que *reactions* foi anotado como um substantivo comum no plural com confiança igual a 1. A ferramenta também pode anotar os sentidos candidatos (quarta coluna) para algumas palavras, por exemplo, entre os sentidos possíveis à palavra *reactions*, a Freeling lista o *synset* com `id = 00859001-n`[3] extraído do Repositório Central Multilíngue da WordNet[4].

Tabela 9 – Exemplos de componentes textuais anotados.

Star_Wars	star_wars	NP00V00 1	
theforceawakens	theforceawakens	VBZ 0.9967	
reactions	reaction	NNS 1	00859001-n:0.00812116...

5.1.4 Consolidação de anotações

Uma palavra pode ter muitas anotações alternativas, inclusive com significados incorretos para o contexto. Além disso, alguns recursos nas bases de conhecimento, como glosa e exemplos de uso, podem estar vazios, ou o significado da palavra não existir na rede semântica utilizada, tais como a denotação de um sentido figurado para as palavras encontrada em alguns *tweets*.

O processo de consolidação de anotações alternativas calcula os níveis de confiança das anotações com base em níveis de confiança retornados pelas ferramentas de anotação (e.g., DBpedia Spotlight, Freeling). Nesta etapa, também são aplicadas regras (e.g. expressões regulares) para normalizar a classificação morfossintática e os sentidos anotados (e.g. granularidade). O resultado de tal consolidação é organizado em descrições de componentes morfossintáticos como definido na Seção 4.1. Para agilizar e evitar problemas na tarefas de desambiguação, pode-se realizar uma poda a fim de reduzir o volume de anotações semânticas de palavras.

[2] `http://nlp.lsi.upc.edu/freeling`
[3] `http://wordnet-rdf.princeton.edu/pwn30/00859001-n`
[4] `http://adimen.si.ehu.es/web/MCR/`

5.1.5 Extração de MSCs

Após as etapas de anotação e consolidação, ao qual as palavras relevantes são anotadas com classes morfossintáticas (marcação de *PoS*) e recursos semânticos (e.g. *synsets* de bases léxicas, recursos LOD) que melhor correspondem ao contexto textual em que elas aparecem, a abordagem MSC^+ extrai os componentes morfo-semânticos. A extração de MSC é realizada de acordo com confiança das anotações geradas pelas ferramentas. A Tabela 10 apresenta os MSCs extraídos da sentença anotada apresentada na Tabela 9. Cada componente possui identificadores *idc* e *idd*, uma palavra arbitrária *w* e *lema*, uma lista de classes morfossintáticas candidatas *P* e uma lista de sentidos candidatos S, respectivamente. Em particular, a abordagem MSC usa dois tipos de características do contexto de palavras: semântica e de linguagem, descritas na sequência.

Tabela 10 – Exemplo de MSC extraído de um *tweet*

	MSC					
#Palavra	*idd*	*idc*	*b*	*e*	*P*	*S*
Star_Wars	1	1	11	19	$\{\langle NPV, 1\rangle\}$	–
theforceawakens	1	2	22	36	$\{\langle VBZ, 0.9966\rangle, \ldots\}$	–
reactions	1	3	44	52	$\{\langle NNS, 1\rangle\}$	$\{\langle 859001\text{-}n\ 0.0081\rangle, \ldots\}$

Características baseadas em semântica. Para cada palavra é extraída uma lista de seus sentidos candidatos com seus respectivos níveis de confiança. Essas informações são obtidas usando ferramentas de anotação que implementam NER/NED e/ou WSI/WSD, geralmente com base em contextos textuais locais. As relações entre *synsets* também podem ser usadas para a marcação dos sentidos. Por exemplo, na WordNet, o *synset* anotado pelo ID `859001-n` (sentido candidato a *idc* 3, coluna S na Tabela 10), que pertence ao grupo morfológico substantivo (*n*), domínio semântico *act*, definido pelo *synset {reaction, response}* que significa "*um processo corporal que ocorre devido ao efeito de algum estímulo ou agente antecedente*" está relacionado aos conjuntos de *synsets* `02894436-a` (*{sensorimotor}*) e `00717358-v` (*{react, respond}*), através das relações "has_derived" e "retated _to", respectivamente.

Características baseadas na linguagem. Este recurso é com base nas probabilidades de n-grama. Em particular, o contexto de uma palavra w é representado por (i) uma ou mais palavras vizinhas à sua esquerda e à sua direita (por exemplo, "*theforceawakens reaction*", "*star_wars reaction*", "*star_wars theforceawakens reaction*"); e (ii) suas respectivas listas de pares candidatos $\langle P, S\rangle$ da classe morfossintática P e sentido S, $\{\langle VBZ, -\rangle, \langle NNS, 859001\text{-}n\rangle\}$, $\{\langle NPV, -\rangle, \langle NNS, 859001\text{-}n\rangle\}$ e $\{\langle NPV, -\rangle, \langle VBZ, -\rangle, \langle NNS, 859001\text{-}n\rangle\}$. As probabilidades de n-grama também são usadas como valores de crença para identificação e detecção de candidatos à marcação de PoS.

5.1.6 Mineração de padrões MSC^+

O algoritmo 1 descreve o método de mineração de padrões MSC^+ e associa a cada um dos padrões descobertos às suas respectivas instâncias correspondentes (perfeitas ou parciais). O algoritmo 1 assume como entradas os componentes morfo-semânticos (MSCs) extraídos de um conjunto de documentos de texto e um suporte mínimo (número de instâncias) $sup \geq 1$ para os padrões retornados.

O parâmetro de suporte do algoritmo 1 foi definido em 2% e 5% em relação ao número de *tweets* de treinamento, sem a necessidade de ajustes adicionais de parâmetro. Após a execução do algoritmo, foram selecionados os padrões de comprimento 3 para a etapa seguinte e avaliação.

A Tabela 11 resume os resultados (número de padrões MSC^+ de comprimento entre 1 e 5 e suas instâncias de correspondência parcial e perfeita) obtidos pela aplicação do algoritmo 1 aos MSCs extraídos do subconjunto selecionado dos *tweets* do Microposts2016, com o suporte definido em 2%. O número de instâncias perfeitamente correspondentes a padrões MSC^+ é maior que o número de instâncias parcialmente correspondentes. Esta situação geralmente ocorre quando o número de MSCs fortemente desambiguados (desconsiderando as palavras sem sentidos anotados pela ferramenta de anotação) é maior que o número de MSCs fracamente desambiguados.

Tabela 11 – Padrões MSC^+ e instâncias encontradas no dataset ($sup = 2\%$).

Comprimento	Padrões MSC^+	Instâncias correspondentes	
		Perfeita	Parcial
1	108	22437	2675
2	1064	57882	14453
3	1943	45512	12726
4	562	5260	1474
5	9	47	9

5.1.7 WSI/WSD com padrões MSC^+

Visando aproveitar a informação semântica capturada pelos padrões MSC^+, foram propostos dois métodos indução/desambiguação lexical de sentido de palavras. Estes métodos combinam os valores de confiança das instâncias fortemente desambiguadas, obtidos de etapas anteriores (anotação morfossintática e de sentidos candidatos e consolidação das anotações), para induzir/desambiguar o sentido de palavras em instâncias fracamente desambiguadas.

Embora essa definição permita parcialmente corresponder instâncias a qualquer número de componentes do MSC com sentido não resolvido, os experimentos relatados nesta tese consideraram apenas instâncias de correspondência parcial às que apresentam apenas um componente de sentido indefinido.

A Tabela 10 ilustra um exemplo de correspondência parcial na sequência de MSCs com dois componentes com o sentido indefinido, representado pelas três palavras componentes "*star_wars*", "*theforceawakens*" e "*reactions*". As duas primeiras palavras não possuem sentidos anotados (identificado pela – na coluna S). Neste caso, cada palavra com o sentido indefinido ("*star_wars*" e "*theforceawekens*"), pertence a uma instância de correspondência parcial e podem ser consideradas como palavras alvo para as tarefas WSI e WSD. Dessa forma, a palavra que está representada por um MSC de sentido indefinido é considerada uma palavras alvo para executar WSI/WSD e anotar um sentido.

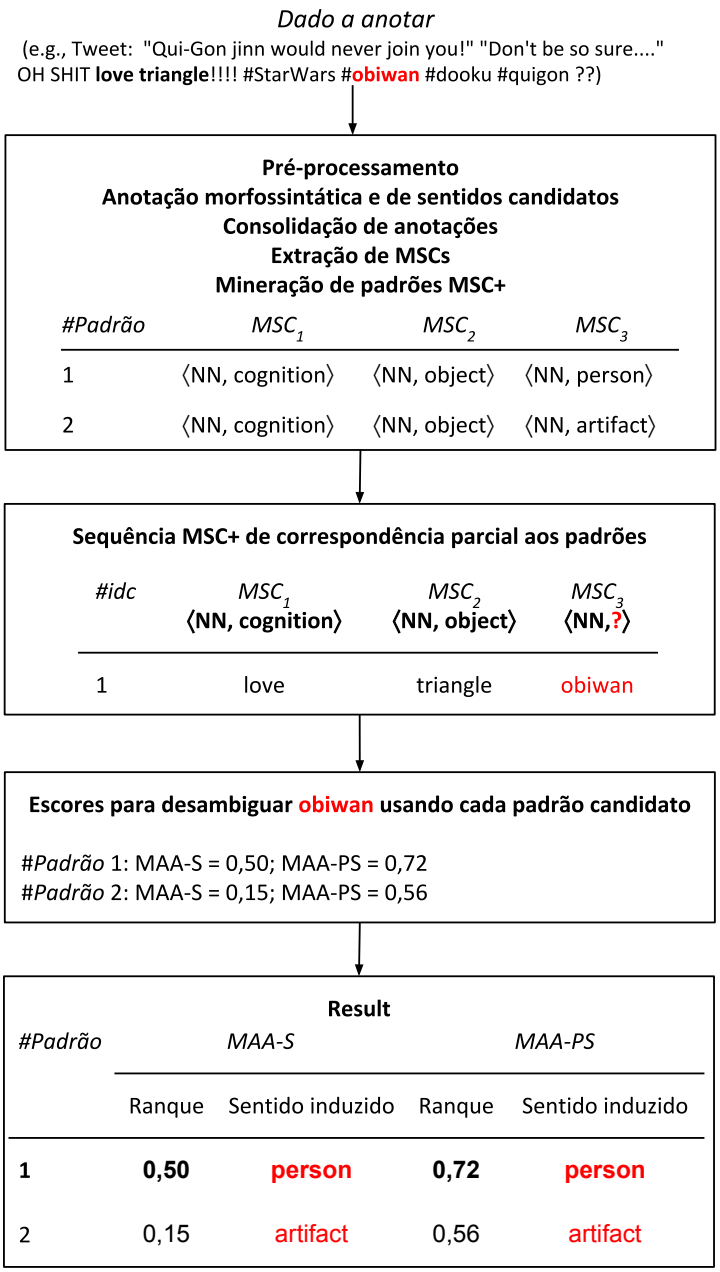
A abordagem MSC^+ para induzir/desambiguar o sentido de palavras consiste em minerar os padrões MSC^+ a partir de um conjunto de documentos de texto, identificando instâncias de padrões que tenham pelo menos um componente com o sentido não definido. Os componentes de sentido indefinido e com correspondência parcial (*matching*) aos padrões minerados, são induzidos para os sentidos anotados nas instâncias de correspondência perfeita, e posteriormente desambiguados. O WSI e o WSD são feitos usando informações heurísticas e estatísticas dos padrões, com suporte suficiente de instâncias de sequências MSC^+ de correspondência perfeita.

A Figura 12 ilustra a aplicação da etapa WSI/WSD (processo mostrado na Figura 11) a um tweet com a sequência MSC (MSC_1, MSC_2, MSC_3), considerando dois padrões MSC^+ extraídos ($\#Padrão$ 1 e 2). Primeiro, o *tweet* é pré-processado para eliminar ruídos e as palavras relevantes são anotadas com as classes morfossintáticas e sentidos candidatos (ferramentas de anotação), após isso, cada MSC relevante é extraído do *tweet* e submetido ao algoritmo 1 para encontrar correspondências entre as sequências MSC^+ do *tweet* e os padrões MSC^+ previamente minerados. No segundo processo da Figura 12 são apresentadas as palavras da sequência MSC (MSC_1, MSC_2, MSC_3), com a palavra alvo $obiwan$ (cujo sentido não foi resolvido pelas ferramentas de anotação) em vermelho na última posição da sequência (MSC_3). Esta sequência MSC^+ corresponde parcialmente aos padrões 1 e 2. O sentido da palavra alvo $obiwan$ (fracamente desambiguado em MSC_3) é induzido pela posição correspondente ao padrão 1 (*person* = pessoa), pois este padrão recebeu a maior pontuação correspondente à posição em MSC_3 para ambos os métodos (MAA-S 0,50 e MAA-PS 0,72).

A Figura 13 demonstra detalhes do cálculo do valor de escore onde são usadas as crenças dos MSCs fortemente desambiguados na posição correspondente (MSC_3) de cada de correspondência perfeita MSC^+ dos padrões candidatos. O método MAA-S considera a média das crenças no sentido ($\tilde{s}$), enquanto que o método MAA-PS considera os dois valores de crenças: o do sentido ($\tilde{s}$, em azul) e o da classe morfossintática ($\tilde{p}$, em verde).

Esta seção descreveu a infraestrutura necessária em todas as etapas utiliza-

Figura 12 – Cálculo do escore com os métodos MAA-S e MAA-PS usando instâncias de correspondência perfeita a cada padrão candidato.



das para induzir/desambiguar o sentido de palavras por meio da abordagem MSC^+, demonstrando detalhadamente o que acontece com os dados até o final do processo. Os resultados referentes ao WSI/WSD com os padrões minerados são relatados e discutidos na Seção 5.2.

Figura 13 – Cálculo dos escores com os métodos MAA-S e MAA-PS usando instâncias de correspondência perfeita a cada padrão candidato.

#Padrão 1:
MAA-S: (0,20 + 1,00 + 0,30) / 3 = 0,50
MAA-PS: (0,99 + 0,86 + 1,00 + 0,20 + 1,00 + 0,30) / (2 * 3) = 0,72

#Padrão 2:
MAA-S: (0,15 + 0,15) / 2 = 0,15
MAA-PS: (0,99 + 0,96 + 0,15 + 0,15) / (2 * 2) = 0,56

Crenças das instâncias de correspondência perfeita

#idc	*#Padrão*	MSC_1 $\langle \tilde{p}_1 , \tilde{s}_1 \rangle$	MSC_2 $\langle \tilde{p}_2 , \tilde{s}_2 \rangle$	MSC_3 $\langle \tilde{p}_3 , \tilde{s}_3 \rangle$
2	1	⟨1,00 , 0,25⟩	⟨0,98 , 0,50⟩	⟨0,99 , 0,20⟩
3	1	⟨1,00 , 0,25⟩	⟨0,98 , 0,50⟩	⟨0,86 , 1,00⟩
4	1	⟨0,98 , 0,08⟩	⟨0,98 , 0,08⟩	⟨1,00 , 0,30⟩
5	2	⟨1,00 , 0,25⟩	⟨0,98 , 0,50⟩	⟨0,99, 0,15⟩
6	2	⟨1,00 , 0,25⟩	⟨0,98 , 0,50⟩	⟨0,96 , 0,15⟩

5.2 RESULTADOS

O dataset Microposts2016 possui um padrão-ouro para entidades que são correspondentes à hierarquia do domínio semântico de palavras da WordNet[5] (também chamado de supersentidos). Os supersentidos são rótulos semânticos de maior granularidade, com base em categorias sintáticas (substantivos, verbos, adjetivos e advérbios) e agrupamentos lógicos de sentidos em classes, tais como pessoa, fenômeno, sentimento e localização (45 grupos no total) (VU *et al.*, 2017). Para medir o desempenho dos métodos MAA-S e MAA-PS na tarefa de indução/desambiguação do sentido foram utilizados os supersentidos, na forma de *um supersentido por colocação* (análogo a um sentido por colocação), e então mapeados para o padrão-ouro do Microposts2016. Para avaliar o desempenho da abordagem WSI/WSD proposta, foram aplicadas as medidas clássicas de precisão, cobertura e medida-F (F_1) tendo em conta a correspondência das classes da DBpedia com as classes de supersentidos retornadas pelas ferramentas de anotação.

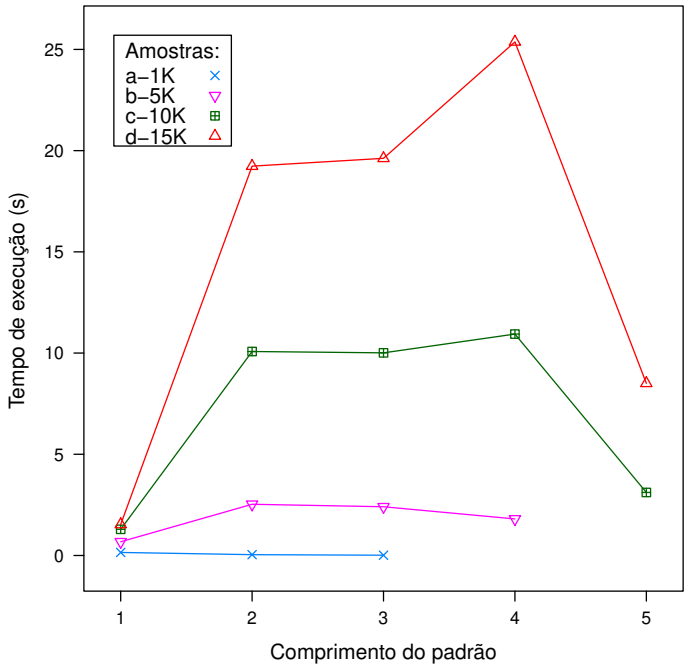
5.2.1 Tempo de execução

Para medir o desempenho do algoritmo 1 em relação ao tempo para minerar os padrões MSC^+ foi usado um conjunto de 20.000 MSCs extraídos dos dados de treina-

[5] `https://wordnet.princeton.edu/documentation/lexnames5wn`

mento do Microposts2016. O conjunto de MSCs foi particionado em quatro amostras: 1.000 MSC (a-1K), 5.000 MSC (b-5K), 10.000 MSC (c-10K) e 15.000 MSC (d-15K).

Figura 14 – Tempo de execução das amostras.



A Figura 14 apresenta a média aritmética do tempo (medido em segundos) necessário em cinco execuções do Algoritmo 1 com cada amostra dos dados. As medições foram feitas usando uma área de trabalho com um processador Intel Core 2 Duo a 3,00 GHz, executando Windows 10, 64 bits e rotinas chamadas via *prompt* em linha de comando.

À medida que o tamanho das amostras aumenta, o cálculo necessário para ser realizado pelo algoritmo 1 para minerar os padrões MSC^+ e combinar as suas instâncias pode acabar sendo substancial, em função da complexidade $O(n^2 \, log \, n)$. No entanto, pode-se realizar particionamentos simples (por exemplo, extração de frases importantes) para não comprometer o tempo de execução do algoritmo quando o tamanho dos dados de entrada for grande, ou ainda, ajustar o parâmetro de suporte.

5.2.2 Indução/Desambiguação do sentido

A anotação do dataset de *tweets* usando FreeLing&Babelfy permitiu extrair 16260 relevantes $MSCs$ fortemente desambiguados, uma vez que 2,6 mil palavras extraídas foram fracamente desambiguadas ou sem nenhum sentido candidato anotado pela Babelfy.

A Tabela 12 apresenta o desempenho WSI/WSD das ferramentas FreeLing (F), FreeLing&Babelfy (FB) (combinadas) e algumas variações quando aplicados os métodos propostos com base nos padrões MSC^+ (abaixo da linha tracejada). As Tabelas 13 e Tabela 14, por sua vez, detalham os ganhos obtidos aplicando por meio da abordagem proposta com o valor de suporte ajustado em 2% e 5%, respectivamente. Cada tabela apresenta o número total de palavras em que o sentido foi anotado pela respectiva abordagem (#Palavras), o número de correspondências em termos de nome de superfície (Menção) e nome de superfície e sentido (Menção&sentido), a cobertura (C), a precisão (P) e a medida-F (F). As colunas C, P e F, nas Tabelas 13 e Tabela 14 referem-se aos ganhos obtidos pela variação dos métodos propostos, que anota os sentidos de várias palavras que estavam com o sentido indefinido (sem nenhum sentido anotado) pelos sistemas de linha de base.

Tabela 12 – Resultados da Indução/Desambiguação do sentido

Sistema	#Palavras	Correspondência		C	P	F
		Menção	Menção&sentido			
FreeLing (F)	13594	586	174	0.709	0.210	0.324
FreeLing&Babelfy (FB)	13770	589	197	0.712	0.238	0.357
F+MAA-PS (MSC2%)	14924	660	179	**0.798****	0.216	0.341
F+MAA-PS (MSC5%)	14226	613	187	0.741	0.225	0.345
FB+MAA-PS (MSC2%)	14455	618	200	0.747*	0.242	**0.365**
FB+MAA-PS (MSC5%)	14046	596	201	0.721	**0.243**	0.364

**p-value < 0.0001, *p-value < 0.05

Tabela 13 – Resultados com padrões MSC^+ e $sup = 2\%$ (MSC 2%)

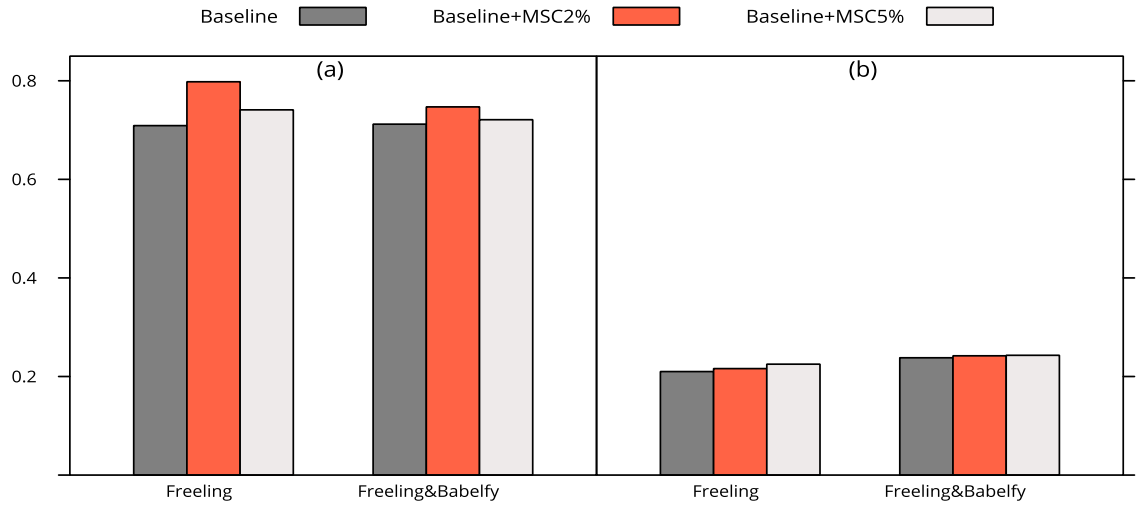
Sistema	Método	#Palavras	Correspondência		C	P	F
			Menção	Menção&sentido			
FreeLing	MAA-S	1330	74	4	0.089	0.005	0.009
	MAA-PS	1330	74	5	0.089	0.006	0.011
FreeLing&	MAA-S	685	29	1	0.035	0.001	0.002
Babelfy	MAA-PS	685	29	3	0.035	0.004	0.007

Tabela 14 – Resultado com padrões MSC^+ e $sup = 5\%$ (MSC 5%)

Sistema	Método	#Palavras	Correspondência		C	P	F
			Menção	Menção&sentido			
FreeLing	MAA-S	632	27	11	0.033	0.013	0.019
	MAA-PS	632	27	12	0.033	0.015	0.020
FreeLing&	MAA-S	227	7	1	0.080	0.001	0.002
Babelfy	MAA-PS	227	7	4	0.080	0.005	0.006

A Tabela 12 destaca em negrito os melhores valores obtidos de precisão, cobertura e medida-F. Os padrões com suporte de 2% (MSC 2%) alcançaram os melhores

Figura 15 – Comparativo entre Cobertura (a) e Precisão (b).



resultados para cobertura, melhorando as anotações geradas pela FreeLing em $8,9\%$ (de $0,709$ para $0,798$) e as geradas na combinação FreeLing&Babelfy em $3,5\%$ (de $0,712$ para $0,747$). Além disso, as variações do método MAA-PS sempre alcançaram um desempenho superior às linhas de base correspondentes. A significância estatística destes ganhos foi demonstrada conforme a Tabela 15 comprovada aplicando o teste Binomial (considerando $\alpha = 0,05$). Os ganhos significativos foram encontrados apenas para a cobertura, com o valor de p $< 0,0001$ (**) para o maior valor em relação a Freeling, e p $< 0,05$ (*) para o maior valor em relação a FreeLing&Babelfy.

Tabela 15 – Teste de hipóteses sobre a significância dos ganhos com o método MAA-PS (experimentos do tipo Binomial).

		Valor-p para C		*Valor-p* para P		*Valor-p* para F	
Sistema	n	MSC2%	MSC5%	MSC2%	MSC5%	MSC2%	MSC5%
FreeLing	827	0,0000	0,0724	0,3828	0,2298	0,2315	0,1827
FreeLing&Babelfy	827	0,0544	0,3423	0,4245	0,4060	0,3674	0,3834

O método proposto MAA-PS obteve os melhores resultados de precisão nos sistemas com os suporte testados (MSC 2% e MSC 5%). No sistema FreeLing, a precisão de $0,210$ e a medida-F de $0,324$ puderam ser melhoradas para $0,225$ e $0,345$, respectivamente, usando FreeLing+MAA-PS (MSC 5%). No sistema FreeLing&Babefy, a precisão de $0,238$ e a medida-F de $0,357$ foram aprimoradas para $0,243$ e $0,364$, respectivamente, usando FreeLing&Babelfy+MAA-PS (MSC 5%). O sistema FreeLing&Babefy+MAA-PS (MSC 2%) alcançou o melhor resultado geral usando na abordagem MSC^+.

A Tabela 12 também mostra o desempenho em relação às palavras alvos (#Palavras) candidatas à tarefa de desambiguação. Usando os sistemas FreeLing e Free-

Ling&Babelfy com o suporte MSC 2%, os resultados mostraram um ganho superior em relação ao MSC 5% (Tabela 13), 1330 palavras-alvo de destino para FreeLing e 685 para FreeLing&Babelfy, respectivamente. A mineração dos padrões MSC^+ com um valor de suporte mais baixo (MSC 2%) permite encontrar mais palavras alvos candidatas porque os sequências MSC^+ sofrem poucas podas. O fato mencionado também está relacionado com a quantidade de anotações geradas pelas ferramentas, e que quando combinadas, tendem a apresentar mais palavras anotadas, consequentemente diminuindo o número de palavras alvo candidatas.

A Figura 15 compara graficamente os resultados de cada sistema de linha de base com os as combinações usando o MAA-PS da abordagem MSC^+, método que obteve o melhor desempenho em termos de cobertura (a) e precisão (b). Considerando o MSC 2% (suporte = 2%), o método MAA-PS encontra mais instâncias candidatas para a tarefa de desambiguação do que quando o valor de suporte é ajustado para 5% (MSC 5%), assim alcançando os melhores resultados de cobertura dentre os sistemas testados. Por outro lado, o MSC 5% obteve melhores resultados de precisão em relação ao MSC 2%. Isto ocorre porque, se aumentarmos o valor de suporte, as sequências MSC correspondentes precisam ocorrer várias vezes no conjunto de dados para serem consideradas um padrão e, consequentemente, o Algoritmo 1 encontrará um número menor de padrões.

O MAA-S utiliza apenas a crença dos sentidos das palavras de instâncias fortemente desambiguadas, enquanto que o MAA-SP utiliza a combinação das crenças de marcações de PoS e sentidos das palavras. Em ambos resultados, na Tabela 13 e na Tabela 14, o método MAA-SP permitiu alcançar a melhor precisão nos dois sistemas testados com os diferentes valores de suporte.

5.3 CONSIDERAÇÕES

O experimento realizado demonstrou que a classificação morfossintática e as anotações semânticas baseadas em inventários de sentidos e bases de conhecimento podem ser usadas para descobrir padrões de linguagem em texto de mídia social. Os padrões minerados fornecem informações de contexto mais gerais que complementam as de documentos específicos, assim como melhoram o desempenho da tarefa WSD em *tweets*. O experimento também demonstra que a confiança em anotações distintas (ou seja, classes morfossintáticas candidatas e sentidos das palavras) são úteis para melhorar a precisão e a cobertura da WSD.

Uma possibilidade em trabalhos futuros é explorar a abordagem proposta e os métodos para induzir/desambiguar o sentido de palavras em experimentos mais extensivos, utilizando outros conjuntos de dados e sistemas existentes. A abordagem proposta pode ser aplicada facilmente com outras ferramentas diferentes da FreeLing e Babelfy, além de inventários com sentido de palavras e bases de conhecimento

semânticas diferentes do WordNet e BabelNet.

O algoritmo proposto para minerar os padrões MSC^+, assim como os *datasets* usados nos experimentos estão disponíveis no Github[6].

[6] https://github.com/fabiobif/MSC-patterns

6 USO DE PADRÕES MSC^+ EM ANÁLISE MORFO-SEMÂNTICA DE DISCURSOS

A semântica, na perspectiva da linguística, estuda o significado das palavras, frases e textos de uma língua natural (JURAFSKY; MARTIN, 2009). Uma técnica que visa a compreensão da língua natural e do seu uso é a Análise do Discurso (AD) ou Análise Automática do Discurso (AAD), para a qual o texto é produto da atividade discursiva que representa aspectos de contexto histórico-social de seus autores (PÊCHEUX *et al.*, 1990).

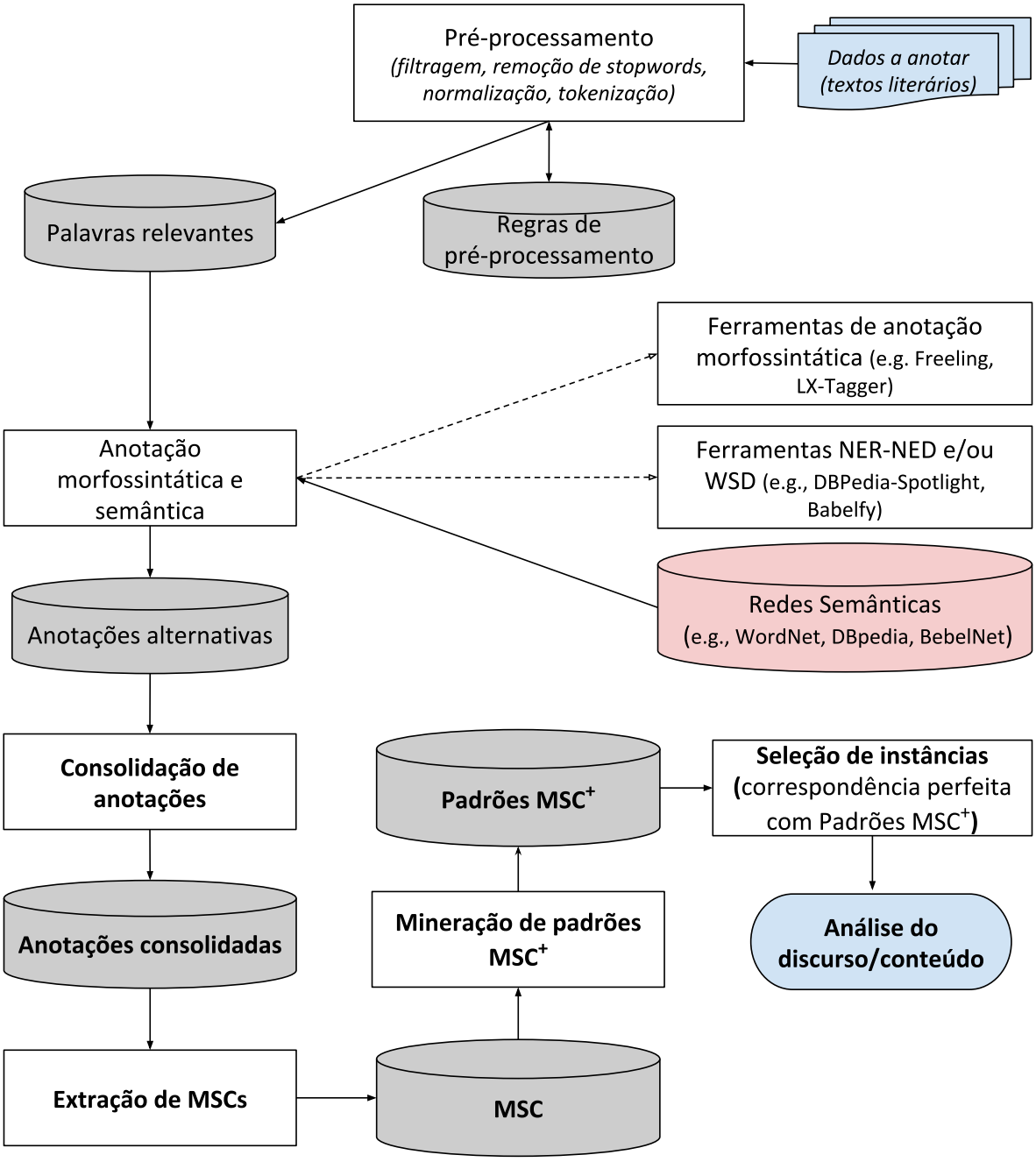
A AD não é uma atividade trivial, pois dependendo das construções linguísticas empregadas e de fatores como a ambiguidade, que pode gerar diferenças e até contradições no modo como certas composições léxicas e sintáticas são interpretadas. Por exemplo, existem textos que apresentam uma semântica desafiadora ou contexto limitado, como acontece com algumas obras literárias e em mídias sociais (ANTONIOU; VAN HARMELEN, 2004). Nesses casos, a eficiência na análise automática do discurso pode ser comprometida, assim como os processos de ensino e aprendizado que se utilizam desses recursos. Considerando os desafios que a semântica de obras literárias impõe à prática da AD, este experimento apresenta uma abordagem computacional para apoiar a AAD por meio de padrões MSC^+, com um estudo de caso em análise de textos literários.

Este capítulo apresenta o experimento realizado com a abordagem MSC^+ para investigar a utilização de instâncias de padrões MSC^+ com casamento perfeito e de alta confiança para a análise do discurso/conteúdo. A Seção 6.1 descreve a infraestrutura do experimento e as ferramentas utilizadas. A maior parte das etapas que compõem o experimento 6.1 são análogas às etapas do experimento anterior, em desambiguação lexical de sentido. Sendo assim, as subseções seguintes detalham, principalmente, os resultados de cada etapa, sem aprofundamento em conceitos que já foram explicados em seções anteriores.

6.1 DESCRIÇÃO DO EXPERIMENTO

Coleções de textos, como obras literárias, tendem a apresentar hábitos de linguagem semelhantes e que podem fornecer informações contextuais. O experimento aqui realizado tira proveito desse fato para extrair padrões da linguagem que são comuns em um coleção de texto e utilizá-los em AAD. A Figura 16 ilustra as principais fases da abordagem proposta para AD, que é composta por seis etapas: (i) `Pré-processamento`, (ii) `Anotação morfossintática e semântica`, (iii) `Consolidação de anotações`, (iv) `Extração de padrões MSCs`, (v) `Mineração de padrões` MSC^+ e (vi) `Instâncias de correspondência perfeita com padrões` MSC^+.

Figura 16 – Abordagem proposta para AAD.



6.1.1 Base de dados utilizada

A abordagem proposta foi aplicada à obras selecionadas de *Machado de Assis*[1] com gêneros literários distintos: *"Contos Fluminenses"* (**Conto**), *"Ressurreição"* (**Romance**) e *"Crisálidas"* (**Poesias**). A Tabela 16 apresenta alguns exemplos do texto encontrado nas obras deste conjunto de dados.

[1] As primeiras obras do autor na Biblioteca Digital de Literaturas de Língua Portuguesa (BDLP) `https://www.literaturabrasileira.ufsc.br`.

Tabela 16 – Uma amostra de textos do conjunto de dados.

"Contos Fluminenses" **(Conto)**: $\simeq 55052$ palavras
Era conveniente ao romance que o leitor ficasse muito tempo sem saber quem era Miss Dollar. Mas por outro lado, sem a apresentação de Miss Dollar, seria o autor obrigado a longas digressões, que encheriam o papel sem adiantar a ação. Não há hesitação possível: vou apresentar-lhes Miss Dollar...

"Crisálidas" **(Poesia)**: $\simeq 12887$ palavras
Lúcia O Dilúvio Fé A Caridade A Jovem Cativa No Limiar Aspiração Cleópatra Os Arlequins As Ondinas Maria Duplessis As Rosas Os DOUS Horizontes Monte Alverne As Ventoinhas Alpujarra Embirração Posfácio...

"Ressurreição" **(Romance)**: 36908 palavras
Não sei o que deva pensar deste livro; ignoro sobretudo o que pensará dele o leitor. A benevolência com que foi recebido um volume de contos e novelas, que há dois anos publiquei, me animou a escrevê-lo. É um ensaio. Vai despretensiosamente às mãos da crítica e do público, que o tratarão com a justiça que merecer...

6.1.2 Pré-processamento

Primeiramente, é necessário submeter o texto à tarefa de pré-processamento, para segmentar parágrafos em sentenças menores e, realizar um limpeza, que remove as partes menos úteis do texto, como pontuação e palavras irrelevantes (*stopwords*). Da mesma maneira que executado no experimento 1, foram consideradas as partes úteis do texto como sendo *tokens* que adicionam informações relevantes para a mensagem como **verbos**, **substantivos**, **advérbios** e **adjetivos**. A Tabela 17 ilustra o trecho de exemplo da obra *"Ressurreição"* **(Romance)** de entrada e como ele é transformado após o estágio de pré-processamento (saída).

Tabela 17 – Um exemplo de pré-processamento de texto.

Entrada:
Não sei o que deva pensar deste livro; ignoro sobretudo o que pensará dele o leitor. A benevolência com que foi recebido um volume de contos e novelas, que há dois anos publiquei, me animou a escrevê-lo. É um ensaio. Vai despretensiosamente às mãos da crítica e do público, que o tratarão com a justiça que merecer.

Saída:
sei deva pensar deste livro ignoro sobretudo pensará leitor
benevolência recebido volume contos novelas dois anos publiquei animou escrevê-lo
Vai despretensiosamente mãos crítica público tratarão justiça merecer

6.1.3 Anotação morfossintática e semântica

A fase de anotação morfossintática e semântica tem por objetivo enriquecer os componentes do texto, resultantes da fase de pré-processamento. Nesta fase, dife-

rentes ferramentas de anotação morfossintática (e.g. FreeLing[2], LxParser[3]) e bases de conhecimento (e.g., WordNet[4], BabelNet[5]) com possíveis valores de anotações semânticas podem ser utilizadas.

A título ilustrativo a Tabela 18 apresenta anotações geradas pela FreeLing em componentes textuais resultantes do pré-processamento do texto *"A frente da casa dava para uma rua pouco frequentada e quase sem moradores"*. A ferramenta FreeLing associa a cada *token* seu lema, classe morfossintática com um valor de confiança, e uma lista de possíveis sentidos que são representados por *synsets* da WordNet.

Tabela 18 – Exemplo de anotações de componentes em texto literário

#	Freeling				WordNet	
	Token	Lema	PoS	Confiança	Sentido	Confiança
1	frente	frente	NC	1,00	08573674-n, ...	0,85, ...
2	casa	casa	NC	1,00	03528100-n, ...	0,90 , ...
3	dava	dar	VMI	0,99	01629000-v, ...	0,79 , ...
4	rua	rua	NC	1,00	04335209-n, ...	0,91 , ...
5	pouco	pouco	RG	1,00	00100002-r, ...	0,99 , ...
6	frequentada	frequentar	VMP	0,99	01844431-v	1,00
7	quase	quase	RG	1,00	-	-
8	moradores	morador	NC	1,00	10523519-n, ...	0,90 , ...

Nota: NC - Nome comum, VMI - Verbo principal modo indicativo, RG - Advérbio geral, VMP - Verbo principal modo particípio

A abordagem proposta usa dois tipos de características do contexto textual: (a) **semântico** - para cada palavra existe uma lista de sentidos candidatos com seus respectivos níveis de confiança. Esta informação é obtida usando ferramentas de anotação que implementam reconhecimento e desambiguação de entidades (NER/NED) e/ou WSD; e (b) **contexto da linguagem** - características baseados em probabilidades de n-grama[6]. Por exemplo, o contexto de uma palavra *w* é representado por uma ou mais palavras vizinhas à esquerda e à direita. As probabilidades de n-grama também são usadas como valores de confiança para os candidatos à marcação de PoS.

Como resultado do pré-processamento (Seção 6.1.2) das obras selecionadas foram extraídas as sentenças com três ou mais palavras relevantes para as etapas de anotação e mineração. Tais sentenças foram submetidas ao processo de anotação usando a ferramenta FreeLing e a base de conhecimento léxico WordNet. O número

[2] http://nlp.lsi.upc.edu/freeling/
[3] http://lxcenter.di.fc.ul.pt/services/pt/LXParserPT.html
[4] http://wordnet.pt/
[5] https://babelnet.org/
[6] Uma sequência contígua de n itens de uma amostra de texto

de sentenças anotadas de **Conto**, **Romance** e **Poesia** totalizaram 3753, 2346 e 569, respectivamente.

6.1.4 Consolidação de anotações

O processo de consolidação de anotações utilizado neste experimento é idêntico ao realizado no experimento com WSD. Basicamente, os valores de confiança das anotações são calculados com base nos níveis de confiança retornados pelas ferramentas de anotação Freeling e BabelNet, e algumas regras de normalização são aplicadas por meio de expressões regulares.

6.1.5 Extração de MSCs

O conteúdo de textos literários pode ter seus significados captados por meio de padrões morfo-semânticos (Seção 4.1). Um MSC refere-se a uma palavra de um documento que possui relação com um conjunto de candidatas.

A extração de MSC é feita com base nos valores de confiança das anotações feitas por ferramentas de última geração. Os MSCs podem aparecer em um documento de texto separados uns dos outros por palavras consideradas irrelevantes. Uma sequência de msc (**instância de padrão** MSC^+) é definida como uma lista ordenada de MSC, que pode estar fracamente ou fortemente desambigua.

6.1.6 Mineração de padrões MSC^+ e seleção de instâncias

Nesta etapa, o algoritmo 1 minera os padrões MSC$^+$ e associa a cada um deles as suas respectivas instâncias correspondentes no texto. As instâncias com correspondência dita **perfeita**, ocorrem quando todos os componentes são fortemente desambiguados (alta confiança) e casam perfeitamente em termos de classe morfossintática e sentido $\langle p, s \rangle$. As instâncias com correspondência dita **parcial**, ocorrem quando ao menos um dos componentes tem o seu sentido indefinido (e.g. Tabela 18 #7), não desambiguado, ou apenas fracamente desambiguado (baixa confiança).

O algoritmo 1 usa como entrada os MSCs extraídos do conjunto de documentos de texto literários e um valor de suporte mínimo (número de instâncias) para retornar os padrões minerados. Os MSCs nas sentenças são automaticamente extraídos e da mesma forma que o experimento 1, são rotulados no formato de dados vertical.

As sentenças anotadas permitiram a mineração de 28440, 19481 e 7126 padrões MSC^+ nas obras. O parâmetro de suporte do algoritmo de mineração foi fixado em 1% do número de sentenças produzidas para o treinamento. Nenhum ajuste de parâmetro adicional foi necessário.

A Tabela 19 apresenta a quantidade de padrões MSC^+ de cada tamanho que foram minerados (coluna MSC^+), com as respectivas quantidades de instâncias (co-

lunas Perfeito e Parcial), nas obras de cada estilo (super-colunas **Conto**, **Romance** e **Poesias**). Selecionamos os padrões de maior tamanho (3) encontrados em todas as obras, pois supostamente carregam o máximo sem ser demasiadamente extensos, para avaliar seu uso em AD na próxima subseção.

Tabela 19 – Quantidades de padrões minerados e instâncias nos documentos

Tam.	**Conto**			**Romance**			**Poesias**		
	MSC$^+$	Perfeita	Parcial	MSC^+	Perfeita	Parcial	MSC^+	Perfeita	Parcial
1	43	28552	9296	43	20150	5974	38	6396	1944
2	156	21253	39408	203	21223	27769	362	25801	20080
3	4	88	10638	10	331	17456	331	17456	49527
4	-	-	-	-	-	-	40	1789	35222

6.2 RESULTADOS

A abordagem proposta consiste em minerar os padrões MSC^+ de um conjunto de documentos de texto, associar as respectivas instâncias (casamento parcial ou perfeito) aos padrões, e por fim, utilizar as instâncias com casamento perfeito e de alta confiança para a análise do discurso/conteúdo.

A análise do discurso/conteúdo é feita usando informações estatísticas dos padrões com confiança suficiente para corresponder ao casamento perfeito entre classe morfossintática e sentido às instâncias de sequência do MSC^+. É justamente o mapeamento de repetições de casamento perfeito em quantidade significativa que permite estabelecer (em análises literárias) padrões estilísticos de escritores. No caso de uma limitação mais restritiva dos padrões de repetição, podem-se quantificar fronteiras que, a partir de um patamar mínimo de proximidade, configurariam situações de citação quase direta ou até mesmo de plágio.

Os métodos de AD são direcionados à análise de certos elementos do discurso, tais como os aspectos sociais, a sua forma e encadeamento, ou ainda, as relações entre seus elementos constitutivos (estrutura). A análise das relações entre os elementos que definem a semântica implícita no discurso (e.g. relação entre classe morfossintática[7] e sentido[8]) foi realizada investigando-se a distribuição das anotações geradas pela ferramenta utilizada nos experimentos (Figura 17) e pela análise dos padrões MSC$^+$ (Tabela 20). De uma forma geral, os documentos analisados apresentaram a distribuição de classe morfossintática e sentido similar, apenas variando em relação à frequência. A Figura 17 ilustra tal fato observado, assim como as classes predominantes: Substantivo (SUB) e Verbo (VER).

[7] `https://talp-upc.gitbook.io/freeling-4-0-user-manual/tagsets`
[8] `https://wordnet.princeton.edu/documentation/lexnames5wn`

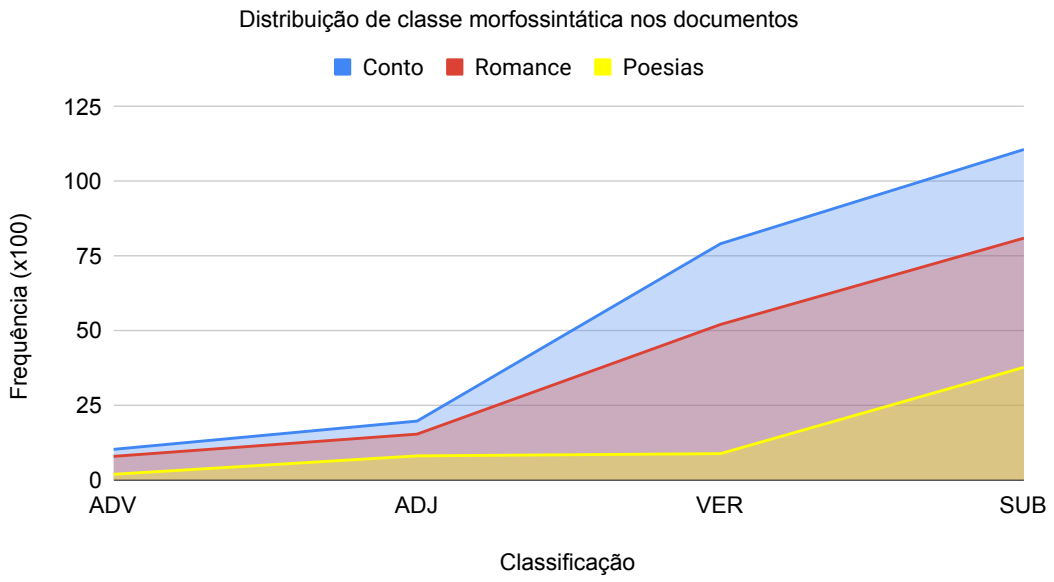
Com base nos documentos analisados são feitas as seguintes considerações sobre o discurso:

1. **Conto** - é composto por narrativas sobre histórias com moral, exemplos que segundo o autor não devem ser seguidos. Elas transmitem a mensagem de personagens malfeitores que ao final são punidos. Também atribui papel de destaque a personagens femininas, assim como humor e sentimento de compaixão.

2. **Romance** - apresenta o discurso sobre a história de um homem solteiro de 36 anos que, apesar de não acreditar no amor, se apaixona por uma viúva. Tal homem apresenta-se emocionalmente instável, desconfiado, inseguro e com sentimentos de ciúmes, o que resulta na separação definitiva do casal. Além disso, o discurso enfatiza a descrição da natureza e dos costumes históricos de uma época.

3. **Poesias** - várias delas foram dedicadas a uma mulher. O discurso neste documento possui um significado mais romântico, sem dispensar o próprio diálogo com textos bíblicos.

Figura 17 – Prevalência das anotações de classe morfossintática nos documentos



A Tabela 20 apresenta uma síntese dos padrões MSC$^+$ de maior prevalência (N) nos documentos, juntamente com exemplos prominentes de sequências de palavras (instâncias) aderentes a cada um. Em relação ao documento **Conto**, os padrões MSC$^+$ de maior prevalência combinam verbo (*VMI*) e interjeição (*I*), ambos com o sentido de processos comunicativos (e.g., responder, dizer, contar). A obra original da qual

Tabela 20 – Os três padrões MSC$^+$ mais frequentes no discurso

Tipo Doc.	**Padrão MSC$^+$** $\langle PoS^*, Sentido^{**}\rangle$	**N**	**Instâncias mais comuns**
Conto	$\langle I, com\rangle\langle VMI, com\rangle\langle I, com\rangle$	22	responder evitar evitar, dizer enganar contar, tirar fingir mostrar, ...
	$\langle VMI, com\rangle\langle I, com\rangle\langle VMI, com\rangle$	22	contar descobrir ler, dizer contar arranjar, concluir concluir perdoar, ...
	$\langle VMI, com\rangle\langle VMI, com\rangle\langle I, com\rangle$	22	voltar comunicar perguntar, emendar reparar ofender, enganar contar arranjar, ...
Romance	$\langle NCF, cond\rangle\langle NCF, perc\rangle\langle NCF, cond\rangle$	47	alegria impressão vida, aliança família situação, vida influência separação, ...
	$\langle NCF, cond\rangle\langle NCF, atr\rangle\langle NCF, cond\rangle$	48	vida herança pobreza, relação pessoa confiança, vez injustiça suspeita, ...
	$\langle NCF, perc\rangle\langle NCF, perc\rangle\langle NCF, cond\rangle$	77	aparência reputação realidade, razão loucura felicidade, opinião moléstia melancolia, ...
Poesias	$\langle AQC, all\rangle\langle NCF, cond\rangle\langle NCF, atr\rangle$	120	doce noite hora, flamejante tristeza beleza, solene solidão natureza, ...
	$\langle NCF, atr\rangle\langle AQF, all\rangle\langle NCF, cond\rangle$	128	calma indiferente visão, tristeza melhor fantasia, natureza rude vida, ...
	$\langle NCF, cond\rangle\langle AQF, all\rangle\langle NCF, cond\rangle$	150	glória universal dor, fé comum dor, liberdade celeste aura, ...

* I - Interjeição, VMI - Verbo principal modo indicativo, NCF - Substantivo comum feminino, AQC - Adjetivo quantitativo comum, AQF - Adjetivo quantitativo feminino.
** com - Processo e conteúdo comunicativo, per - Processo cognitivo, cond - Condição particular de alguém ou algo, atr - Atributo de pessoas e objetos, all - Geral.

se extraiu o documento **Conto** possui 359 interjeições (!) apontadas em seu discurso, enquanto as demais, **Romance** e **Poesias**, 107, 138, respectivamente. Os principais padrões MSC$^+$ encontrados no documento **Romance** são formados pela combinação de substantivo feminino (*NCF*), com os sentidos relacionados à condições específicas de uma pessoa (e.g., alegria, aliança, confiança) e a processos cognitivos (e.g., impressão, aparência, reputação). Em **Poesias**, os padrões encontrados são formados na maioria por substantivo feminino (*NCF*) com o sentido de condição/situação (e.g., liberdade, dor) e atributo (e.g., hora, calma, beleza), combinado a adjetivo feminino (*AQF*) de sentido geral.

6.3 CONSIDERAÇÕES

Os padrões textuais explorados neste trabalho (MSC$^+$) levam em conta a ordem de ocorrência das palavras com base em marcação de PoS e classes de sentidos, o que permite considerar o conteúdo semântico relacional (e.g. relações específicas que os nomes e os verbos mantêm entre si) e não apenas a mera correspondência lexical. Os experimentos reportados revelam que a classificação morfossintática aliada à anotação semântica permitem encontrar padrões linguísticos que podem auxiliar na AAD de conteúdos literários. Isso fica evidente desde a observação da preponderância de verbos nos contos, de substantivos nos romances e de adjetivos (ou de substantivos derivados destes) nos poemas.

Os diferentes padrões MSC$^+$ minerados mostram as diferenças entre os gêneros literários, o que pode ser elemento importante e clarificador em atividades de ensino-aprendizagem de literatura. Além disso, alguns resultados fornecem evidências científicas e estatísticas para subsidiar pesquisas sobre os possíveis usos da abordagem MSC^+, não só em AAD, mas também na resolução do significado de palavras de textos literários.

Os experimentos relatados neste estudo são prontamente reproduzíveis, pois o algoritmo para mineração de padrões MSC^+ e as obras literárias utilizadas estão publicamente disponíveis no Github[9]. Como trabalhos futuros pretende-se avaliar a capacidade da abordagem proposta na melhoria de anotações semânticas (precisão e cobertura) sobre os resultados da ferramenta de anotação, explorar métricas que possam avaliar o conteúdo captado pelos padrões MSC^+ no discurso das obras, assim como definir estratégias para multiplicar as perspectivas de leitura do objeto literário.

[9] https://github.com/fabiobif/MSC-patterns

7 CONCLUSÃO

Neste capítulo, as conclusões obtidas acerca do problema de mineração e uso de padrões linguísticos (MSC^+) para a desambiguação de palavras e análise do discurso são apresentadas e analisadas. Primeiramente, as hipóteses de pesquisa e objetivos previamente definidos no Capítulo 1 são recuperados e discutidos segundo os resultados obtidos nesta tese. Estas conclusões são realizadas com base em experimentos realizados que correspondem às principais contribuições deste trabalho. Na sequência são apresentadas as principais publicações que auxiliam a corroborar as contribuições apresentadas neste capítulo. Por fim, também são discutidas as limitações da abordagem proposta, evidenciando sua efetividade em diferentes cenários de aplicação, assim como os trabalhos e pesquisas futuras.

7.1 PRINCIPAIS CONTRIBUIÇÕES

Esta tese apresentou uma abordagem para explorar padrões de linguagem compostos por sequências de componentes morfo-semânticos (MSC), bastante frequentes em alguns documentos de texto curtos, como *tweets*. A abordagem proposta utiliza um algoritmo para minerar esses padrões, que foram denominados de padrões MSC^+, a partir de conjuntos de documentos de texto. Os padrões MSC^+ se baseiam nas anotações de classes morfossintáticas e sentidos de palavras relevantes. As anotações podem ser produzidas por uma variedade de ferramentas alternativas disponíveis para PNL.

O algoritmo para minerar os padrões MSC^+ pode ser visto como um tipo de aprendizado automático de padrões linguísticos que captam as informações contextuais. Os padrões podem ser usados para resolver o sentido das palavras que, de outra forma, seriam erroneamente rotuladas, ou ainda para a análise do discurso/conteúdo de texto, conforme apresentado nos experimentos (Capítulos 5 e 6). Os métodos propostos para induzir/desambiguar o sentido de palavras baseados nos padrões MSC^+ podem complementar as abordagens atuais para a resolução da semântica lexical, tanto em uma perspectiva conceitual quanto empírica, mesmo quando as informações de contexto são bastante limitadas em documentos específicos.

A abordagem proposta é uma alternativa que explora um tipo específico de padrão de linguagem (MSC^+) utilizando ferramentas consideradas como tradicionais (e.g., classificadores morfossintáticos, anotadores semânticos, bases e redes de conhecimento léxico) em seu processo. Os resultados obtidos, ainda que de maneira modesta, demonstraram que o aprendizado de padrões usado na abordagem MSC^+ auxilia ferramentas de classificação na melhoria de anotações semânticas em textos, empregando métodos que exigem menos processamento quando comparados à técnicas que envolvem aprendizado profundo (Deep learning). Em publicações de outros

pesquisadores, padrões análogos aos definidos nesta tese podem ser obtidos usando incorporação de palavras (Word embedding) produzidas por técnicas de aprendizado profundo, e usados para análise do discurso e classificação de texto, além de WSI e WSD. Portanto, o aprendizado utilizado na abordagem MSC^+ pode ser combinado ao aprendizado profundo na tarefa de WSI/WSD porque são consideradas abordagens complementares, em vez de concorrentes.

A abordagem MSC^+ permitem analisar um texto de forma superficial, com eficiência e alta cobertura, usando ferramentas que beneficiam tarefas de extração e recuperação de informação, ao contrário da análise de texto profunda.

7.2 CONTRIBUIÇÕES CIENTÍFICAS

As contribuições científicas desta tese estão relacionadas às questões de pesquisa (Seção 3.1), hipóteses (Seção 1.2) e objetivos (Seção 1.3) apresentados nesta tese. As questões de pesquisa estão relacionadas aos resultados obtidos com o levantamento sistemático da literatura, enquanto que as hipóteses e objetivos são discutidos em relação aos experimentos realizados para avaliar a abordagem MSC^+ nas tarefas WSD e análise do discurso. Com base nas questões de pesquisa, pode-se extrair as seguintes conclusões sobre o trabalho desenvolvido:

Questão principal e secundárias. *Q1. Com qual frequência e como padrões linguísticos são abordados em WSD? Q1.1 Quais são os principais domínios de aplicação? Q1.2 Quais são os idiomas frequentemente considerados em experimentos? Q1.3 Quais são as fontes de conhecimento/BCL comumente utilizadas? Q1.4 Qual é o processo que prevalece - desambiguação e/ou indução)? Q1.5 Quais técnicas e algoritmos são frequentemente empregados? Q1.6 Quais são as principais abordagens utilizadas?*

Com o objetivo de investigar os trabalhos científicos realizados na área de pesquisa sobre a semântica dos textos, em especial sobre a desambiguação de palavras, foi conduzida uma revisão sistemática da literatura por meio de um plano de pesquisa que define um protocolo de busca e seleção de trabalhos. Várias bibliotecas digitais de pesquisa da área da computação foram acessadas para buscar os trabalhos relevantes sobre o tema. A aplicação da string de busca nas fontes de pesquisa permitiu recuperar 756 estudos, dos quais 135 trabalhos foram selecionados conforme os critérios de exclusão e qualidade estabelecidos. Os estudos selecionados das fontes de pesquisa são correspondentes a momentos distintos, o primeiro em setembro de 2017 e o segundo, que corresponde a uma atualização, em fevereiro de 2019 (inserindo trabalhos publicados até 2018).

A revisão sistemática possibilitou responder às questões de pesquisa com clareza e apontar algumas características importantes sobre a área. Em relação ao do-

mínio de aplicação (Q1.1), de maneira geral existe o predomínio da área de saúde e ciências biológicas (e.g., bioinformática, biomedicina) com a utilização de dados da área médica (e.g., MEDLINE, PubMed), mas com alguns trabalhos que não abrangem o tema desta tese, ainda que relacionados a desambiguação. Como definido no plano de pesquisa, os estudos selecionados foram os que contribuíram de forma efetiva para o estado da arte na tarefa de WSD e sobre o papel da semântica envolvendo o significado de palavras. Neste contexto, o domínio que está em evidência para avaliação semântica são os conjuntos de dados SenEval e Semseval (textos de notícias). Ao se considerar os idiomas (Q1.2), o inglês é preponderante, mas existem trabalhos com experimentos em vários idiomas, como o chinês, italiano, francês, espanhol, alemão, árabe, ou mesmo bengali. Alguns estudos utilizam dados no idioma português, mas este idioma necessita ser mais explorado em pesquisas sobre semântica e desambiguação para ganhar representatividade como o inglês, por exemplo. Sobre as fontes de conhecimento/BCL (Q1.3), WordNet é a mais utilizada, assim como dicionários da web. Contudo, em estudos atuais destacam-se as bases de conhecimento BabelNet, YAGO, SUMO, OntoNotes, DBpedia Spotlight. Considerando o tipo de processo (Q1.4), a desambiguação é predominante em relação à indução do sentido. Entre as técnicas e algoritmos (Q1.5), destaca-se o tratamento da semântica latente, a aplicação de algoritmos tradicionais de AM, algoritmo de Lesk e suas variações. Sobre as principais abordagens (Q1.6), destaca-se o tratamento da semântica por meio de grafos. Abordagens mais recentes utilizam a incorporação de palavras (Word embedding) para aprendizado profundo (Deep learning) em textos. O aprendizado de padrões linguísticos é uma abordagem pouco frequente em WSD. Nos estudos em que esta abordagem é utilizada, as principais características observadas no texto são relações sintáticas, marcação de *PoS*, palavras específicas em um contexto e modelos de representação (e.g. BoW).

Baseado na análise dos resultados obtidos acerca do desempenho da abordagem MSC^{+} nas tarefas WSI/WSD e análise do discurso, pode-se relatar as seguintes conclusões acerca das hipóteses levantadas no desenvolvimento desta tese:

Hipótese 1: Técnicas de sumarização de texto podem detectar sequências de postagens de mídias sociais com semântica mais importante. Textos, especialmente textos coloquiais como postagens em mídias sociais, podem ter muitos componentes semanticamente relevantes, mas que precisam ser identificados.

Hipótese 2: É possível explicitar automaticamente a semântica implícita. A significação de palavras não costuma ser única nem estática. As palavras podem ter seu significado modificado dependendo do contexto, deixando de representar apenas a ideia original (básica e objetiva), remetendo-nos a novos sentidos por meio de associações, às vezes, indiretas (semântica implícita). As informações

de contexto do texto, incluindo do autor e de populações com os traços culturais análogos, podem ser utilizadas para desambiguar ou mesmo induzir o sentido de tais palavras, em várias situações em que isso se faz necessário.

Hipótese 3: É possível melhorar a precisão e cobertura de métodos de anotação semântica por meio de padrões linguísticos. Com a utilização de estruturas semânticas de bases de conhecimento, juntamente com informação de contexto, pode-se extrair a semântica implícita de textos de mídias sociais, e assim captar o significado de palavras que métodos de anotação semântica atuais não fazem.

As técnicas de sumarização de texto podem ser adaptadas para auxiliar na identificação de componentes semanticamente relevantes, especialmente em textos que expressam uma linguagem informal, como postagens de mídias sociais. O *dataset* utilizado no experimento 1 (Capítulo 5) desta tese foi extraído de uma coleção de milhões de *tweets* utilizando algumas palavras-chaves e tópicos relacionados a eventos específicos. Esta seleção de palavras permitiu extrair de um grande volume de dados apenas o texto com semântica relevante no contexto. Em texto de mídia social, este processo de identificar partes do texto onde determinadas palavras ocorrem atribuindo uma importância semântica é análogo a utilização de métricas de sumarização de texto na construção de um resumo.

Nas tarefas selecionadas para os experimentos onde se testou a abordagem MSC^+, pode-se evidenciar palavras com o significado modificado em função do contexto. Quando tais palavras são instâncias fracamente desambiguadas de um padrão MSC^+, elas foram identificadas e classificadas, como mostrado no experimento 1 com WSD.

De acordo com os resultados obtidos, a performance da abordagem MSC^+ melhora significativamente a cobertura e, de maneira modesta, a precisão quando considerada a aplicação baseada em WSD. No caso da análise do discurso a abordagem MSC^+ consegue captar traços semânticos no conteúdo que melhoram a o entendimento sobre características das obras de autores. Contudo, o desempenho evidenciado na análise do discurso necessita de métricas para apresentar uma mensuração efetiva.

Com o **objetivo** (principal) de explorar denotações e relações semânticas de palavras relevantes em *tweets* e textos em língua Portuguesa, para determinar o significado de palavras, neste trabalho foi proposta uma abordagem para o aprendizado de padrões linguísticos. A abordagem foi avaliada e explorada em duas tarefas de PLN que combinam a análise de níveis de processamento da língua (morfológico, sintático e semântico) na representação e interpretação do significado em textos.

Em relação ao **objetivo 1** (teórico) referente à descrição e detalhamento do estado da arte sobre técnicas e ferramentas de anotação semântica automática, prin-

cipalmente daquelas para reconhecimento e classificação de palavras relevantes e desambiguação do sentido de palavras, pode-se destacar as seguintes abordagens utilizadas por meio de procedimentos de enriquecimento textual: redes semânticas (e.g. BabelNet) e bases de conhecimento léxico (e.g. WordNet). Acerca do **objetivo 2** (prático) associado à seleção de ferramentas, bases de dados e bases de conhecimento para realizar a identificação de palavras relevantes, classificação morfossintática e anotação semântica, visando auxiliar no processo de captura de sentidos corretos de palavras, destacam-se as ferramentas Freeling e Babelfy com funcionalidades para vários idiomas, incluindo a língua Portuguesa (e.g. análise morfológica, NER/NED, marcação *PoS*, WSD, rotulagem de função semântica, *Entity linking*, etc); o Microposts2016, um dataset de *tweets* com padrão ouro em anotação semântica; a DBLP, uma biblioteca digital que é fonte primária e gratuita de textos literários em versão integral de autores brasileiros e portugueses. Em relação ao **objetivo 3** (teórico) referente ao desenvolvimento das definições sobre Componente Morfo-semântico (MSC), Padrão Morfo-semântico (MSC^+), o algoritmo para a mineração de padrões e métodos para WSD, destacam-se as investigações iniciais e os conceitos construídos usando formalizações matemáticas. Quanto ao **objetivo 4** (aplicacional) associado à implementação da abordagem proposta, multilingue e aplicável a diferentes conjuntos de dados, para captar a semântica implícita de palavras relevantes em sequências de textos curtos, com base em técnicas de sumarização de texto e análise do contexto, pode-se destacar a definição e uso dos Padrões MSC^+. Em relação ao **objetivo 5** (aplicacional) associado à validação da abordagem proposta em experimentos com dados de microblogs, e textos com ênfase em Português do Brasil, pode-se destacar os resultados com os experimentos realizados em WSD e análise do discurso.

Também é importante destacar objetivos secundários aos definidos nesta tese relacionados à implementação e disponibilização de ferramentas e scripts para auxiliar na aplicação da abordagem MSC^+, assim como todo o conteúdo gerado por meio dos experimentos e que estão disponíveis na Internet.

7.3 PUBLICAÇÕES

Durante o desenvolvimento deste trabalho, as contribuições obtidas foram divulgadas por meio de publicação de artigos em periódicos e publicação e apresentação de artigos em conferências. Essas publicações são listadas a seguir, apresentando a relação de cada uma com este trabalho e indicando aquelas que estão indiretamente relacionadas à esta tese em pesquisas estabelecidas com outros autores.

GOULARTE, FÁBIO BIF; SORATO, D.; NASSAR, SILVIA MODESTO; FILETO, RENATO; SAGGION, HORACIO. MSC+: Language pattern learning for word sense induction and disambiguation. Knowledge-Based Systems, 2019. (*Qualis periódi-*

cos A1).
https://www.sciencedirect.com/science/article/pii/S0950705119304253

Neste artigo foi publicada a abordagem MSC^+ e a proposta de minerar padrões linguísticos em texto de mídia social para classificar a semântica implícita de algumas palavras. Este trabalho pode ser considerado a principal publicação desta tese, resultado de uma exploração de outros trabalhos publicados anteriormente, e diretamente relacionado às questões de pesquisa (Q1.1 a Q1.6).

GOULARTE, FÁBIO BIF; NASSAR, SILVIA MODESTO; Fileto, Renato; SAGGION, HORACIO. A text summarization method based on fuzzy rules and applicable to automated assessment. EXPERT SYSTEMS WITH APPLICATIONS, v. 115, p. 264-275, 2019. (*Qualis periódicos A1*).
https://www.sciencedirect.com/science/article/pii/S0957417418304743.

Neste artigo foi publicado o trabalho referente à uma proposta de avaliação automatizada de atividades acadêmicas por meio da sumarização de texto. Este trabalho auxiliou na formação do argumento que define a primeira hipótese desta pesquisa e na resposta da questão Q1.5 e Q1.6.

SORATO, D.; **GOULARTE** F.B.; FILETO, R. SSP: A Linguistic Pattern Mining Approach for Discourse Analysis Applied to Hate Speech Detection. INTERNATIONAL JOURNAL ON ARTIFICIAL INTELLIGENCE TOOLS, 2019. (*Qualis periódicos B1*).
https://dx.doi.org/10.1142/S0218213020400023

Neste artigo foi publicado os resultados da utilização de padrões linguísticos do tipo SSP para a identificação de um discurso específico. Este trabalho está relacionado a esta pesquisa por apresentar padrões análogos aos utilizados nesta tese, mas usando uma abordagem com incorporação de palavras.

FILETO, R.; OLIVEIRA, I. L.; **GOULARTE**, FB; SORATO, D.; WILLRICH, R.; SANTOS, A. L. Questões e Perspectivas sobre a Produção Automatizada de Anotações Semânticas em Textos Literários In: Simposio de Humanidades Digitales del Sur: Escritura Creativa Digital y Colecciones Digitales, 2017, Madrid. Anais del Simposio de Humanidades Digitales del Sur: Escritura Creativa Digital y Colecciones Digitales, 2017.

Neste artigo foi publicado algumas propostas de trabalhos referentes à exploração da semântica e do significado em texto de mídia social.

SORATO, D.; **GOULARTE**, F. B.; FILETO, R.; NASSAR, S. M. Análise de Métodos e Ferramentas para Reconhecimento de Palavras Relevantes em Microblogs. In: XII Simpósio Brasileiro de Sistemas de Informação, Florianópolis, 2016. (*Qualis*

eventos B2).
https://dl.acm.org/citation.cfm?id=3022014

Neste artigo foi publicado os resultados da utilização de ferramentas de anotação morfossintática na identificação de palavras relevantes em texto de mídia social, em Português. Este trabalho auxiliou na proposição da segunda e terceira hipótese desta pesquisa e na seleção de ferramentas e bases de dados para os experimentos.

Além dos trabalhos publicados apresentados, o seguinte trabalho está em processo de submissão, visando uma publicação no curto prazo.

GOULARTE, D.; NASSAR, S.M.; FILETO, R.; WILLRICH, R.; DOS SANTOS, A. L. Análise de Discurso usando Anotações e Padrões Morfo-Semânticos no Ensino de Literatura. 2020. (Qualis eventos A2, submissão até 01/2020).

7.4 LIMITAÇÕES E TRABALHOS FUTUROS

Uma característica da abordagem MSC^+ refere-se à necessidade de processamento extra dos textos para gerar as representações MSC. O texto utilizado para minerar os padrões MSC^+ deve ser pré-processado e classificado por ferramentas de anotação morfossintática e semântica. Como os textos são curtos (segmentado), pode haver a necessidade de filtragem extra e tarefas adicionais de pré-processamento para ser possível a geração de padrões representativos, pois há limitação na extração de informações semânticas relevantes. Este fato pode ocorrer, por exemplo, quando o texto for um *tweet* com pouco conteúdo ou com poucos termos relevantes para a classificação semântica. Neste sentido, para explorar novas possibilidades de recuperação da informação (*Information Retrieval* - IR) em textos com pouco conteúdo semântico, além da técnica de sumarização de texto, pode-se investigar a aplicação de abordagens de aprendizado profundo, ou então utilizar informação privilegiada (e.g., local e momento da geração do texto, relações semânticas entre as palavras armazenadas nas fontes de conhecimento).

Uma avaliação mais aprofundada da abordagem proposta pode ser realizada usando conjuntos de dados distintos que não envolvam textos de mídia social, como texto breve que pode aparecer em alguns diários e registros médicos. Além disso, pode-se investigar uma variedade de alternativas para melhorar o desempenho da abordagem proposta. Por exemplo, associar pesos globais a cada padrão MSC^+ extraído pelo algoritmo para melhorar o processo de desambiguação. Também é importante investigar os efeitos de várias palavras com sentido não definido na mesma sequência MSC em relação a tarefa WSI/WSD, e maneiras mais inteligentes de incorporar a

influência de palavras fortemente desambiguadas em textos anteriores no suporte e confiança de padrões MSC^+ nas classes morfossintáticas específicas e sentidos.

REFERÊNCIAS

A. JÚNIOR, José Gildo; SCHIEL, Ulrich; MARINHO, Leandro Balby. An approach for building lexical-semantic resources based on heterogeneous information sources. *In:* ACM. PROCEEDINGS of the 30th Annual ACM Symposium on Applied Computing. [*S.l.*: *s.n.*], 2015. p. 402–408.

ABEL, Fabian; GAO, Qi *et al.* Semantic enrichment of twitter posts for user profile construction on the social web. *In:* SPRINGER. EXTENDED Semantic Web Conference. [*S.l.*: *s.n.*], 2011. p. 375–389.

ABEL, Fabian; HAUFF, Claudia *et al.* Semantics+ filtering+ search= twitcident. exploring information in social web streams. *In:* ACM. PROCEEDINGS of the 23rd ACM conference on Hypertext and social media. [*S.l.*: *s.n.*], 2012. p. 285–294.

ABNEY, Steven P. Parsing by chunks. *In:* PRINCIPLE-BASED parsing. [*S.l.*]: Springer, 1991. p. 257–278.

ABOWD, Gregory D *et al.* Towards a better understanding of context and context-awareness. *In:* SPRINGER. INTERNATIONAL Symposium on Handheld and Ubiquitous Computing. [*S.l.*: *s.n.*], 1999. p. 304–307.

AGGARWAL, Charu C; ZHAI, ChengXiang. **Mining text data**. [*S.l.*]: Springer Science & Business Media, 2012.

AGIRRE, Eneko; EDMONDS, Philip. **Word sense disambiguation: Algorithms and applications**. [*S.l.*]: Springer Science & Business Media, 2007. v. 33.

AGIRRE, Eneko; MARTINEZ, David. Knowledge sources for word sense disambiguation. *In:* SPRINGER. TEXT, Speech and Dialogue. [*S.l.*: *s.n.*], 2001. p. 1–10.

AGIRRE, Eneko; MARTINEZ, David *et al.* Two graph-based algorithms for state-of-the-art WSD. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2006 Conference on Empirical Methods in Natural Language Processing. [*S.l.*: *s.n.*], 2006. p. 585–593.

AGIRRE, Eneko; SOROA, Aitor. Personalizing pagerank for word sense disambiguation. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 12th Conference of the European Chapter of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 2009. p. 33–41.

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan *et al.* Mining sequential patterns. *In:* ICDE. [*S.l.*: *s.n.*], 1995. p. 3–14.

AICHNER, T.; JACOB, F. Measuring the degree of corporate social media use. **International Journal of Market Research**, v. 58, n. 2, p. 257–275, 2015.

ALAGIĆ, Domagoj; ŠNAJDER, Jan; PADÓ, Sebastian. Leveraging Lexical Substitutes for Unsupervised Word Sense Induction. *In:* THIRTY-SECOND AAAI Conference on Artificial Intelligence (AAAI-18). [*S.l.*: *s.n.*], 2018.

ALBANO, Lorenzo; BENEVENTANO, Domenico; BERGAMASCHI, Sonia. Word sense induction with multilingual features representation. *In:* IEEE COMPUTER SOCIETY. PROCEEDINGS of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02. [*S.l.*: *s.n.*], 2014. p. 343–349.

AMSLER, Robert A. Lexical knowledge bases. *In:* 10TH International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 1984.

ANANTHARAM, Pramod *et al.* Extracting city traffic events from social streams. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM, v. 6, n. 4, p. 43, 2015.

ANTONIOU, Grigoris; VAN HARMELEN, Frank. **A semantic web primer**. [*S.l.*]: MIT press, 2004.

ANTUNES, João. **Exploração de informações contextuais para enriquecimento semântico em representações de textos**. 2018. Diss. (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em: `http://www.teses.usp.br/teses/disponiveis/55/55134/tde-03012019-103253`.

ARRAIS, Telmo Correia. Estrutura profunda e padrões de representação semântica. **Trans/Form/Ação**, SciELO Brasil, v. 2, p. 138–154, 1975.

ARTILES, Javier; AMIGÓ, Enrique; GONZALO, Julio. The role of named entities in web people search. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. [*S.l.*: *s.n.*], 2009. p. 534–542.

ATTARDI, Giuseppe; DELLORLETTA, Felice. Chunking and dependency parsing. *In:* LREC Workshop on Partial Parsing: Between Chunking and Deep Parsing. [*S.l.*: *s.n.*], 2008.

BAKER, Collin F; FILLMORE, Charles J; LOWE, John B. The berkeley framenet project. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. [*S.l.*: *s.n.*], 1998. p. 86–90.

BALLY, Charles; SECHEHAYE, Albert. **Course in general linguistics Ferdinand de Saussure**. [*S.l.*]: McGraw-Hill Book Company, 1966.

BARONI, Marco *et al.* The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. **Language resources and evaluation**, Springer, v. 43, n. 3, p. 209–226, 2009.

BASHAR, Md Abul *et al.* Conceptual annotation of text patterns. **Computational Intelligence**, Wiley Online Library, v. 33, n. 4, p. 948–979, 2017.

BATEMAN, John; ZOCK, Michael. Natural language generation. *In:* THE Oxford Handbook of Computational Linguistics 2nd edition. [*S.l.*: *s.n.*], 2016.

BÉCHET, Nicolas *et al.* Discovering linguistic patterns using sequence mining. *In:* SPRINGER. INTERNATIONAL Conference on Intelligent Text Processing and Computational Linguistics. [*S.l.*: *s.n.*], 2012. p. 154–165.

BELAININE, Billal; FONSECA, Alexsandro; SADAT, Fatiha. Named Entity Recognition and Hashtag Decomposition to Improve the Classification of Tweets. *In:* PROCEEDINGS of the 2nd Workshop on Noisy User-generated Text (WNUT). [*S.l.*: *s.n.*], 2016. p. 102–111.

BIEMANN, Chris. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the first workshop on graph based methods for natural language processing. [*S.l.*: *s.n.*], 2006. p. 73–80.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural language processing with Python**. [*S.l.*]: "O'Reilly Media, Inc.", 2009.

BONTCHEVA, Kalina; DERCZYNSKI, Leon. Extracting Information from Social Media with GATE. **Working with Text: Tools, Techniques and Approaches for Text Mining**, Elsevier, p. 133, 2016.

BONTCHEVA, Kalina; DERCZYNSKI, Leon *et al.* TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. *In:* RANLP. [*S.l.*: *s.n.*], 2013. p. 83–90.

BONTCHEVA, Kalina; ROUT, Dominic. Making sense of social media streams through semantics: a survey. **Semantic Web**, IOS Press, v. 5, n. 5, p. 373–403, 2014.

BRILL, Eric. Part-of-speech tagging. **Handbook of natural language processing**, Marcel Dekker, p. 403–414, 2000.

BRODY, Samuel; LAPATA, Mirella. Bayesian word sense induction. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 12th Conference of

the European Chapter of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 2009. p. 103–111.

CABEZUDO, Marco Antonio Sobrevilla; PARDO, Thiago Alexandre Salgueiro *et al.* Exploratory study of word sense disambiguation methods for verbs in brazilian portuguese. **International Journal of Computational Linguistics and Applications**, Bahri Publications, v. 6, n. 1, p. 131–148, 2015.

CAI, Jun Fu; LEE, Wee Sun; TEH, Yee Whye. NUS-ML: Improving word sense disambiguation using topic features. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 4th International Workshop on Semantic Evaluations. [*S.l.*: *s.n.*], 2007. p. 249–252.

CAMACHO-COLLADOS, José; PILEHVAR, Mohammad Taher; NAVIGLI, Roberto. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. **Artificial Intelligence**, Elsevier, v. 240, p. 36–64, 2016.

CAMBRIA, Erik; WHITE, Bebo. Jumping NLP curves: a review of natural language processing research [review article]. **IEEE Computational Intelligence Magazine**, IEEE, v. 9, n. 2, p. 48–57, 2014.

CAO, Ziqiang *et al.* TGSum: Build Tweet Guided Multi-Document Summarization Dataset. *In:* AAAI. [*S.l.*: *s.n.*], 2016. p. 2906–2912.

CEWEB. **Guia de Web Semântica**. [*S.l.*: *s.n.*], 2016. http://ceweb.br/guias/web-semantica/. Accessed: aug., 2016.

CHANG, Haw-Shiuan *et al.* Efficient Graph-based Word Sense Induction by Distributional Inclusion Vector Embeddings. **arXiv preprint arXiv:1804.03257**, 2018.

CHANG, Yi *et al.* Timeline Summarization from Social Media with Life Cycle Models. *In:* IJCAI. [*S.l.*: *s.n.*], 2016. p. 3698–3704.

CHEN, Xinxiong; LIU, Zhiyuan; SUN, Maosong. A Unified Model for Word Sense Representation and Disambiguation. *In:* EMNLP. [*S.l.*: *s.n.*], 2014. p. 1025–1035.

CHOMSKY, Noam. **Chomsky, N. 1959. A review of BF Skinner's Verbal behavior. Language, 35 (1), 26–58.** [*S.l.*]: JSTOR, 1959.

CHURCH, Kenneth Ward. A stochastic parts program and noun phrase parser for unrestricted text. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the second conference on Applied natural language processing. [*S.l.*: *s.n.*], 1988. p. 136–143.

COECKE, Bob; SADRZADEH, Mehrnoosh; CLARK, Stephen. Mathematical foundations for a compositional distributional model of meaning. **arXiv preprint arXiv:1003.4394**, 2010.

CRUSE, Alan. **Meaning in Language: An Introduction to Semantics and Pragmatics**. [*S.l.*]: Oxford University Press UK, 2011.

D'AGOSTINI, Caio Stein; FILETO, Renato. Capturing Users' Preferences and Intentions in a Semantic Search System. *In:* PROCEEDINGS of the 21st International Conference on Software Engineering & Knowledge Engineering (SEKE'2009), Boston, Massachusetts, USA, July 1-3, 2009. [*S.l.*]: Knowledge Systems Institute Graduate School, 2009. p. 587–591. ISBN 1-891706-24-1.

DAIBER, Joachim *et al.* Improving efficiency and accuracy in multilingual entity extraction. *In:* ACM. PROCEEDINGS of the 9th International Conference on Semantic Systems. [*S.l.*: *s.n.*], 2013. p. 121–124.

DAS, Amitava; BURMAN, Utsab *et al.* NER from Tweets: SRI-JU System@ MSM 2013. **Making Sense of Microposts (# MSM2013)**, 2013.

DAS, TK; ACHARJYA, DP; PATRA, MR. Opinion mining about a product by analyzing public tweets in Twitter. *In:* IEEE. COMPUTER Communication and Informatics (ICCCI), 2014 International Conference on. [*S.l.*: *s.n.*], 2014. p. 1–4.

DECAPUA, A. **Grammar for Teachers: A Guide to American English for Native and Non-Native Speakers**. [*S.l.*]: Springer International Publishing, 2016. (Springer Texts in Education). ISBN 9783319339160. Disponível em: `https://books.google.com.br/books?id=S9TBDAAAQBAJ`.

DEDEK, Jan. **Semantic Annotations**. 2012. Tese (Doutorado) – Faculty of Mathematics e Physics, Charles University.

DEHAENE, Stanislas *et al.* Conscious, preconscious, and subliminal processing: a testable taxonomy. **Trends in cognitive sciences**, Elsevier, v. 10, n. 5, p. 204–211, 2006.

DERCZYNSKI, Leon; MAYNARD, Diana *et al.* Analysis of named entity recognition and linking for tweets. **Information Processing & Management**, Elsevier, v. 51, n. 2, p. 32–49, 2015.

DERCZYNSKI, Leon; RITTER, Alan *et al.* Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *In:* RANLP. [*S.l.*: *s.n.*], 2013. p. 198–206.

DI MARCO, Antonio; NAVIGLI, Roberto. Clustering and diversifying web search results with graph-based word sense induction. **Computational Linguistics**, MIT Press, v. 39, n. 3, p. 709–754, 2013.

DIAS-DA-SILVA, Bento Carlos; DI FELIPPO, Ariani; NUNES, Maria das Graças Volpe. The Automatic Mapping of Princeton WordNet Lexical-Conceptual Relations onto the Brazilian Portuguese WordNet Database. *In:* LREC. [*S.l.*: *s.n.*], 2008. p. 335–342.

DRAGONI, Mauro. Computational advertising in social networks: an opinion mining-based approach. *In:* ACM. PROCEEDINGS of the 33rd Annual ACM Symposium on Applied Computing. [*S.l.*: *s.n.*], 2018. p. 1798–1804.

EHRMANN, Maud *et al.* Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. *In:* LREC. [*S.l.*: *s.n.*], 2014. p. 401–408.

ELTYEB, Safaa; SALIM, Naomie. Chemical named entities recognition: a review on approaches and applications. **Journal of cheminformatics**, Springer International Publishing, v. 6, n. 1, p. 17, 2014.

FELDMAN, Ronen; SANGER, James *et al.* **The text mining handbook: advanced approaches in analyzing unstructured data**. [*S.l.*]: Cambridge university press, 2007.

FELLBAUM, Christiane. **WordNet**. [*S.l.*]: Wiley Online Library, 1998.

FILETO, Renato *et al.* The Baquara 2 knowledge-based framework for semantic enrichment and analysis of movement data. **Data & Knowledge Engineering**, Elsevier, v. 98, p. 104–122, 2015.

FILLMORE, Charles J. Types of lexical information. *In:* STUDIES in syntax and semantics. [*S.l.*]: Springer, 1969. p. 109–137.

FLORIN, J.L. **Introdução à linguıstica: objetos teóricos**. [*S.l.*]: Contexto, 2002. (Introducão á linguıstica). ISBN 9788572441926. Disponível em: `https://books.google.com.br/books?id=JDfsswEACAAJ`.

FODOR, Jerry. Semantics: an interview with Jerry Fodor. **Revista Virtual de Estudios da Linguagem**, v. 5, n. 8, 2007.

FOKKENS, Antske *et al.* Offspring from Reproduction Problems: What Replication Failure Teaches Us. *In:* ACL (1). [*S.l.*: *s.n.*], 2013. p. 1691–1701.

FOURNIER-VIGER, Philippe *et al.* A survey of sequential pattern mining. **Data Science and Pattern Recognition**, v. 1, n. 1, p. 54–77, 2017.

GAONA, Miguel Ángel Rıos; GELBUKH, Alexander; BANDYOPADHYAY, Sivaji. Web-based variant of the lesk approach to word sense disambiguation. *In:* IEEE. ARTIFICIAL Intelligence, 2009. MICAI 2009. Eighth Mexican International Conference on. [*S.l.*: *s.n.*], 2009. p. 103–107.

GATTANI, Abhishek *et al.* Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. **Proceedings of the VLDB Endowment**, VLDB Endowment, v. 6, n. 11, p. 1126–1137, 2013.

GENC, Yegin; SAKAMOTO, Yasuaki; NICKERSON, Jeffrey. Discovering context: classifying tweets through a semantic transform based on wikipedia. **Foundations of augmented cognition. Directing the future of adaptive systems**, Springer, p. 484–492, 2011.

GILSTER, Paul. **Digital literacy**. [*S.l.*]: Wiley Computer Pub., 1997.

GOULARTE, Fábio Bif; NASSAR, Silvia Modesto *et al.* A text summarization method based on fuzzy rules and applicable to automated assessment. **Expert Systems with Applications**, Elsevier, v. 115, p. 264–275, 2019.

GOULARTE, Fábio Bif; SORATO, Danielly *et al.* MSC+: Language pattern learning for word sense induction and disambiguation. **Knowledge-Based Systems**, Elsevier, p. 105017, 2019.

GREGORY, Lorenzo; RAVELLI, Andrea Amelio; PANUNZI, Alessandro. Enriching BabelNet verbal entities with videos: a linking experiment with the IMAGACT ontology of action. *In:* THE Luxembourg BabelNet Workshop. [*S.l.*: *s.n.*], 2016. p. 2–3.

GUREVYCH, Iryna; ECKLE-KOHLER, Judith; MATUSCHEK, Michael. Linked lexical knowledge bases: Foundations and applications. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers, v. 9, n. 3, p. 1–146, 2016.

GUTIÉRREZ, Yoan; VÁZQUEZ, Sonia; MONTOYO, Andres. Spreading semantic information by Word Sense Disambiguation. **Knowledge-Based Systems**, Elsevier, v. 132, p. 47–61, 2017.

HAMMERTON, James *et al.* Introduction to special issue on machine learning approaches to shallow parsing. **Journal of Machine Learning Research**, v. 2, Mar, p. 551–558, 2002.

AL-HARBI, Omar; JUSOH, Shaidah; NORWAWI, Norita Md. Lexical disambiguation in natural language questions (NLQs). **arXiv preprint arXiv:1709.09250**, 2017.

HARRIS, Zellig S. Distributional structure. **Word**, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954.

HE, Ruifang *et al.* Twitter summarization with social-temporal context. **World Wide Web**, Springer, v. 20, n. 2, p. 267–290, 2017.

HEARST, Marti A. Untangling text data mining. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 37th annual meeting of the

Association for Computational Linguistics on Computational Linguistics. [*S.l.*: *s.n.*], 1999. p. 3–10.

HENRICH, Verena. **Word Sense Disambiguation with GermaNet**. 2015. Tese (Doutorado) – Eberhard Karls Universität Tübingen.

HILPINEN, Risto. Conception, sense, and reference in Peircean semiotics. **Synthese**, Springer, v. 192, n. 4, p. 991–1018, 2015.

HUA, Wen; ZHENG, Kai; ZHOU, Xiaofang. Microblog Entity Linking with Social Temporal Context. *In:* ACM. PROCEEDINGS of the 2015 ACM SIGMOD International Conference on Management of Data. [*S.l.*: *s.n.*], 2015. p. 1761–1775.

HUANG, Eric H *et al.* Improving word representations via global context and multiple word prototypes. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. [*S.l.*: *s.n.*], 2012. p. 873–882.

JONES, Bevan *et al.* Semantics-Based Machine Translation with Hyperedge Replacement Grammars. *In:* COLING. [*S.l.*: *s.n.*], 2012. p. 1359–1376.

JUNG, Jason J. Online named entity recognition method for microtexts in social networking services: A case study of twitter. **Expert Systems with Applications**, Elsevier, v. 39, n. 9, p. 8066–8070, 2012.

JURAFSKY, Daniel; MARTIN, James H. Speech and Language processing: An introduction to natural language processing. **Computational Linguistics and Speech Recognition, 2nd Ed., Prentice Hall**, 2009.

JURGENS, David; KLAPAFTIS, Ioannis. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. *In:* SECOND Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). [*S.l.*: *s.n.*], 2013. p. 290–299.

JURGENS, David; PILEHVAR, Mohammad Taher. Reserating the awesometastic: An automatic extension of the WordNet taxonomy for novel terms. *In:* HLT-NAACL. [*S.l.*: *s.n.*], 2015. p. 1459–1465.

KILGARRIFF, Adam *et al.* The Sketch Engine: ten years on. **Lexicography**, Springer, v. 1, n. 1, p. 7–36, 2014.

KIM, Sun; LU, Zhiyong; WILBUR, W John. Identifying named entities from PubMed® for enriching semantic categories. **BMC bioinformatics**, BioMed Central, v. 16, n. 1, p. 57, 2015.

KINGSBURY, Paul; PALMER, Martha. From TreeBank to PropBank. *In:* LREC. [*S.l.*: *s.n.*], 2002. p. 1989–1993.

KITCHENHAM, Barbara; BRERETON, O Pearl *et al.* Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, Elsevier, v. 51, n. 1, p. 7–15, 2009.

KITCHENHAM, Barbara; CHARTERS, Stuart. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007.

KUDO, Taku; MATSUMOTO, Yuji. Chunking with support vector machines. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. [*S.l.*: *s.n.*], 2001. p. 1–8.

KUMAR, Saket. **Word Sense Disambiguation**. 2015. Tese (Doutorado) – University of Pennsylvania.

KUPIEC, Julian. An algorithm for finding noun phrase correspondences in bilingual corpora. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 31st annual meeting on Association for Computational Linguistics. [*S.l.*: *s.n.*], 1993. p. 17–22.

LAENDER, Alberto HF *et al.* A brief survey of web data extraction tools. **ACM Sigmod Record**, ACM, v. 31, n. 2, p. 84–93, 2002.

LANDAUER, Thomas K; FOLTZ, Peter W; LAHAM, Darrell. An introduction to latent semantic analysis. **Discourse processes**, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998.

LANIADO, David; MIKA, Peter. Making sense of twitter. **The Semantic Web–ISWC 2010**, Springer, p. 470–485, 2010.

LAU, Jey Han *et al.* Word sense induction for novel sense detection. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 13th Conference of the European Chapter of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 2012. p. 591–601.

LEACOCK, Claudia; CHODOROW, Martin. Combining local context and WordNet similarity for word sense identification. **WordNet: An electronic lexical database**, v. 49, n. 2, p. 265–283, 1998.

LEHMANN, Jens *et al.* DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. **Semantic Web**, v. 6, n. 2, p. 167–195, 2015. DOI: `10.3233/SW-140134`. Disponível em: `http://dx.doi.org/10.3233/SW-140134`.

LESK, Michael. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *In:* ACM. PROCEEDINGS of the 5th annual international conference on Systems documentation. [*S.l.*: *s.n.*], 1986. p. 24–26.

LEU, Donald J *et al.* Toward a theory of new literacies emerging from the Internet and other information and communication technologies. **Theoretical models and processes of reading**, v. 5, n. 1, p. 1570–1613, 2004.

LI, Jiwei; JURAFSKY, Dan. Do multi-sense embeddings improve natural language understanding? **arXiv preprint arXiv:1506.01070**, 2015.

LIN, Dekang. Automatic retrieval and clustering of similar words. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 17th international conference on Computational linguistics-Volume 2. [*S.l.*: *s.n.*], 1998. p. 768–774.

LIN, Dekang; PANTEL, Patrick. DIRT@ SBT@ discovery of inference rules from text. *In:* ACM. PROCEEDINGS of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. [*S.l.*: *s.n.*], 2001. p. 323–328.

LIN, Dekang; PANTEL, Patrick. Discovery of inference rules for question-answering. **Natural Language Engineering**, Cambridge University Press, v. 7, n. 4, p. 343–360, 2001.

LIN, Thomas; ETZIONI, Oren *et al.* Entity linking at web scale. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction. [*S.l.*: *s.n.*], 2012. p. 84–88.

LINOFF, Gordon S; BERRY, Michael JA. **Data mining techniques: for marketing, sales, and customer relationship management**. [*S.l.*]: John Wiley & Sons, 2011.

LIU, Peng-Yuan *et al.* PengYuan@ PKU: Extracting infrequent sense instance with the same N-gram pattern for the SemEval-2010 task 15. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 5th International Workshop on Semantic Evaluation. [*S.l.*: *s.n.*], 2010. p. 371–374.

LIU, Xiaohua *et al.* Joint inference of named entity recognition and normalization for tweets. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. [*S.l.*: *s.n.*], 2012. p. 526–535.

LUND, Kevin; BURGESS, Curt. Producing high-dimensional semantic spaces from lexical co-occurrence. **Behavior Research Methods, Instruments, & Computers**, v. 28, n. 2, p. 203–208, jun. 1996. ISSN 1532-5970. DOI: `10.3758/BF03204766`. Disponível em: `https://doi.org/10.3758/BF03204766`.

MAHDISOLTANI, Farzaneh; BIEGA, Joanna; SUCHANEK, Fabian. Yago3: A knowledge base from multilingual wikipedias. *In:* CIDR CONFERENCE. 7TH Biennial Conference on Innovative Data Systems Research. [*S.l.*: *s.n.*], 2014.

MAHDISOLTANI, Farzaneh; BIEGA, Joanna; SUCHANEK, Fabian M. Yago3: A knowledge base from multilingual wikipedias. *In:* CIDR. [*S.l.*: *s.n.*], 2013.

MARCUS, Mitchell P; MARCINKIEWICZ, Mary Ann; SANTORINI, Beatrice. Building a large annotated corpus of English: The Penn Treebank. **Computational linguistics**, MIT Press, v. 19, n. 2, p. 313–330, 1993.

MARTIN, Egoitz Laparra. Implicit Semantic Roles in Discourse, 2015.

MEIJ, Edgar; WEERKAMP, Wouter; RIJKE, Maarten de. Adding semantics to microblog posts. *In:* ACM. PROCEEDINGS of the fifth ACM international conference on Web search and data mining. [*S.l.*: *s.n.*], 2012. p. 563–572.

MENDES, Ana Cristina; ANTUNES, Cláudia. Pattern mining with natural language processing: An exploratory approach. *In:* SPRINGER. INTERNATIONAL Workshop on Machine Learning and Data Mining in Pattern Recognition. [*S.l.*: *s.n.*], 2009. p. 266–279.

MENDES, Pablo N; JAKOB, Max *et al.* DBpedia spotlight: shedding light on the web of documents. *In:* ACM. PROCEEDINGS of the 7th international conference on semantic systems. [*S.l.*: *s.n.*], 2011. p. 1–8.

MEYER, Christian M. **Wiktionary: The Metalexicographic and the Natural Language Processing Perspective**. 2013. Tese (Doutorado) – Technische Universität.

MIHALCEA, Rada. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. [*S.l.*: *s.n.*], 2005. p. 411–418.

MIHALCEA, Rada. Using wikipedia for automatic word sense disambiguation. *In:* HLT-NAACL. [*S.l.*: *s.n.*], 2007. p. 196–203.

MIHALCEA, Rada F. Word sense disambiguation with pattern learning and automatic feature selection. **Natural Language Engineering**, Cambridge University Press, v. 8, n. 4, p. 343–358, 2002.

MIKOLOV, Tomas *et al.* Distributed representations of words and phrases and their compositionality. *In:* ADVANCES in neural information processing systems. [*S.l.*: *s.n.*], 2013. p. 3111–3119.

MILLER, George A *et al.* Introduction to WordNet: An on-line lexical database. **International journal of lexicography**, Oxford Univ Press, v. 3, n. 4, p. 235–244, 1990.

MILLER, Tristan. **Adjusting Sense Representations for Word Sense Disambiguation and Automatic Pun Interpretation**. 2016. Tese (Doutorado) – tuprints.

MILLER, Tristan *et al.* Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. *In:* COLING. [*S.l.*: *s.n.*], 2012. p. 1781–1796.

MILLWARD, Celia M; HAYES, Mary. **A biography of the English language**. [*S.l.*]: Nelson Education, 2011.

MONDAL, Anupam *et al.* Lexical resource for medical events: A polarity based approach. *In:* IEEE. 2015 IEEE International Conference on Data Mining Workshop (ICDMW). [*S.l.*: *s.n.*], 2015. p. 1302–1309.

MORGAN, Mena Badieh Habib; VAN KEULEN, Maurice. Information extraction for social media. *In:* PROCEEDINGS of the Third Workshop on Semantic Web and Information Extraction. [*S.l.*: *s.n.*], 2014. p. 9–16.

MORO, Andrea; CECCONI, Francesco; NAVIGLI, Roberto. Multilingual Word Sense Disambiguation and Entity Linking for Everybody. *In:* PROCEEDINGS of the 13th International Semantic Web Conference, Posters and Demonstrations (ISWC 2014). Riva del Garda, Italy: [*s.n.*], 2014. p. 25–28.

MORO, Andrea; NAVIGLI, Roberto. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. *In:* SEMEVAL@ NAACL-HLT. [*S.l.*: *s.n.*], 2015. p. 288–297.

MORO, Andrea; RAGANATO, Alessandro; NAVIGLI, Roberto. Entity linking meets word sense disambiguation: a unified approach. **Transactions of the Association for Computational Linguistics**, v. 2, p. 231–244, 2014.

NAVIGLI, Roberto. A quick tour of word sense disambiguation, induction and related approaches. *In:* SPRINGER. INTERNATIONAL Conference on Current Trends in Theory and Practice of Computer Science. [*S.l.*: *s.n.*], 2012. p. 115–129.

NAVIGLI, Roberto. Word sense disambiguation: A survey. **ACM Computing Surveys (CSUR)**, ACM, v. 41, n. 2, p. 10, 2009.

NAVIGLI, Roberto; LAPATA, Mirella. An experimental study of graph connectivity for unsupervised word sense disambiguation. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 32, n. 4, p. 678–692, 2010.

NAVIGLI, Roberto; PONZETTO, Simone Paolo. BabelNet: Building a very large multilingual semantic network. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 48th annual meeting of the association for computational linguistics. [*S.l.*: *s.n.*], 2010. p. 216–225.

NG, Hwee Tou; LEE, Hian Beng. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 34th annual meeting on Association for Computational Linguistics. [*S.l.*: *s.n.*], 1996. p. 40–47.

NGOMO, Axel-Cyrille Ngonga *et al.* Introduction to Linked Data and Its Lifecycle on the Web. *In:* REASONING on the Web in the Big Data Era, Athens, Greece. [*S.l.*: *s.n.*], 2014. p. 1–99. DOI: `10.1007/978-3-319-10587-1_1`. Disponível em: `http://dx.doi.org/10.1007/978-3-319-10587-1_1`.

NÓBREGA, Fernando Antônio Asevêdo. **Desambiguação lexical de sentidos para o português por meio de uma abordagem multilıngue mono e multidocumento**. 2013. Tese (Doutorado) – Universidade de São Paulo.

NÓBREGA, Fernando Antônio Asevedo; PARDO, Thiago Alexandre Salgueiro. General purpose word sense disambiguation methods for nouns in Portuguese. *In:* SPRINGER. INTERNATIONAL Conference on Computational Processing of the Portuguese Language. [*S.l.*: *s.n.*], 2014. p. 94–101.

NOUVEL, Damien; EHRMANN, Maud; ROSSET, Sophie. **Named Entities for Computational Linguistics**. [*S.l.*]: John Wiley & Sons, 2016.

PAGE, Lawrence *et al.* **The PageRank citation ranking: Bringing order to the web.** [*S.l.*], 1999.

PANCHENKO, Alexander. Best of Both Worlds: Making Word Sense Embeddings Interpretable. Versão inglesa. *In:* CHAIR), Nicoletta Calzolari (Conference *et al.* (Ed.). **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)**. Portoro?, Slovenia: European Language Resources Association (ELRA), mai. 2016.

PANCHENKO, Alexander *et al.* Unsupervised, Knowledge-Free, and Interpretable Word Sense Disambiguation. **arXiv preprint arXiv:1707.06878**, 2017.

PAZZANI, Michael J; BILLSUS, Daniel. Content-based recommendation systems. *In:* THE adaptive web. [*S.l.*]: Springer, 2007. p. 325–341.

PÊCHEUX, Michel *et al.* Análise automática do discurso. **Por uma análise automática do discurso: uma introdução à obra de Michel Pêcheux**, Editora da UNICAMP Campinas, v. 2, p. 61–151, 1990.

PELEVINA, Maria *et al.* Making sense of word embeddings. **arXiv preprint arXiv:1708.03390**, 2017.

PETERSEN, Kai; VAKKALANKA, Sairam; KUZNIARZ, Ludwik. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology**, Elsevier, v. 64, p. 1–18, 2015.

PISKORSKI, Jakub; YANGARBER, Roman. Information extraction: Past, present and future. *In:* MULTI-SOURCE, multilingual information extraction and summarization. [*S.l.*]: Springer, 2013. p. 23–49.

PONZETTO, Simone Paolo; NAVIGLI, Roberto. Knowledge-rich word sense disambiguation rivaling supervised systems. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 48th annual meeting of the association for computational linguistics. [*S.l.*: *s.n.*], 2010. p. 1522–1531.

PONZETTO, Simone Paolo; NAVIGLI, Roberto. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. *In:* IJCAI. [*S.l.*: *s.n.*], 2009. p. 2083–2088.

POSSENTI, Sırio. **Por que (não) ensinar gramática na escola**. [*S.l.*]: Mercado de Letras Campinas, 1996.

QUINIOU, Solen *et al.* What about sequential data mining techniques to identify linguistic patterns for stylistics? *In:* SPRINGER. INTERNATIONAL Conference on Intelligent Text Processing and Computational Linguistics. [*S.l.*: *s.n.*], 2012. p. 166–177.

RAMSHAW, Lance A; MARCUS, Mitchell P. Text chunking using transformation-based learning. *In:* NATURAL language processing using very large corpora. [*S.l.*]: Springer, 1999. p. 157–176.

REN, Zhaochun. **Monitoring social media: Summarization, classification and recommendation**. 2016. Tese (Doutorado) – Universiteit van Amsterdam.

REN, Zhaochun *et al.* Personalized time-aware tweets summarization. *In:* ACM. PROCEEDINGS of the 36th international ACM SIGIR conference on Research and development in information retrieval. [*S.l.*: *s.n.*], 2013. p. 513–522.

RITTER, Alan; CLARK, Sam; ETZIONI, Oren *et al.* Named entity recognition in tweets: an experimental study. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROC. of the Conf. on Empirical Methods in Natural Language Processing. [*S.l.*: *s.n.*], 2011. p. 1524–1534.

RODRIGUEZ, Mayra Z *et al.* Clustering Algorithms: A Comparative Approach. **arXiv preprint arXiv:1612.08388**, 2016.

ROMERO, Simone Aparecida Pinto. **A framework for event classification in Tweets based on hybrid semantic enrichment**. 2017. Tese (Doutorado) – Universidade Federal do Rio Grande do Sul.

ROSENBERG, Andrew; HIRSCHBERG, Julia. V-measure: A conditional entropy-based external cluster evaluation measure. *In:* PROCEEDINGS of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). [*S.l.*: *s.n.*], 2007.

RUDRA, Koustav *et al.* Summarizing situational tweets in crisis scenario. *In:* ACM. PROCEEDINGS of the 27th ACM Conference on Hypertext and Social Media. [*S.l.*: *s.n.*], 2016. p. 137–147.

RUIZ-CASADO, Maria; ALFONSECA, Enrique; CASTELLS, Pablo. Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. **Data & Knowledge Engineering**, Elsevier, v. 61, n. 3, p. 484–499, 2007.

SACENTI, Juarez AP *et al.* Automatically tailoring semantics-enabled dimensions for movement data warehouses. *In:* SPRINGER. INTERNATIONAL Conference on Big Data Analytics and Knowledge Discovery. [*S.l.*: *s.n.*], 2015. p. 205–216.

SAGGION, Horacio; POIBEAU, Thierry. Automatic text summarization: Past, present and future. *In:* MULTI-SOURCE, multilingual information extraction and summarization. [*S.l.*]: Springer, 2013. p. 3–21.

SAGGION, Horacio; ŠTAJNER, Sanja *et al.* Making it simplext: Implementation and evaluation of a text simplification system for spanish. **ACM Transactions on Accessible Computing (TACCESS)**, ACM, v. 6, n. 4, p. 14, 2015.

SALTON, Gerard; WONG, Anita; YANG, Chung-Shu. A vector space model for automatic indexing. **Communications of the ACM**, ACM, v. 18, n. 11, p. 613–620, 1975.

SANTOS, Fábio; GONÇALO OLIVEIRA, Hugo. Descoberta de Synsets Difusos com base na Redundância em vários Dicionários. **Linguamática**, v. 7, n. 2, p. 3–17, 2015.

SCARTON, Carolina. **VerbNet. Br: construção semiautomática de um léxico verbal online e independente de domınio para o português do Brasil. NILC/USP**. 2013. Tese (Doutorado) – Dissertação de mestrado orientada por Sandra Maria Aluísio.

SCARTON, Carolina; DURAN, Magali Sanches; ALUÍSIO, Sandra Maria. Using Cross-Linguistic Knowledge to Build VerbNet-Style Lexicons: Results for a (Brazilian) Portuguese VerbNet. *In:* SPRINGER. INTERNATIONAL Conference on Computational Processing of the Portuguese Language. [*S.l.*: *s.n.*], 2014. p. 149–160.

SCHINAS, Manos *et al.* MGraph: multimodal event summarization in social media using topic models and graph-based ranking. **International Journal of Multimedia Information Retrieval**, Springer, v. 5, n. 1, p. 51–69, 2016.

SCHNEIDER, Nathan *et al.* Semeval-2016 task 10: Detecting minimal semantic units and their meanings (dimsum). *In:* PROCEEDINGS of the 10th International Workshop on Semantic Evaluation (SemEval-2016). [*S.l.*: *s.n.*], 2016. p. 546–559.

SCHULER, Karin Kipper. **VerbNet: A broad-coverage, comprehensive verb lexicon**. 2005. Tese (Doutorado) – University of Pennsylvania.

SCHULZ, Axel; GUCKELSBERGER, Christian; JANSSEN, Frederik. Semantic Abstraction for generalization of tweet classification: An evaluation of incident-related tweets. **Semantic Web**, IOS Press, v. 8, n. 3, p. 353–372, 2017.

SCHÜTZE, Hinrich. Automatic word sense discrimination. **Computational linguistics**, MIT Press, v. 24, n. 1, p. 97–123, 1998.

SEBASTIANI, Fabrizio. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM, v. 34, n. 1, p. 1–47, 2002.

SHARIFI, Beaux; HUTTON, Mark-Anthony; KALITA, Jugal. Summarizing microblogs automatically. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. HUMAN Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 2010. p. 685–688.

SHEN, Wei; WANG, Jianyong; HAN, Jiawei. Entity linking with a knowledge base: Issues, techniques, and solutions. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 27, n. 2, p. 443–460, 2015.

SIDDHARTHAN, Advaith; MANDYA, Angrosh. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. *In:* EACL. [*S.l.*: *s.n.*], 2014. p. 722–731.

SILVA, Alberto Rodrigues da. Linguistic Patterns and Linguistic Styles for Requirements Specification (I) An Application Case with the Rigorous RSL/Business-Level Language. *In:* PROCEEDINGS of the 22nd European Conference on Pattern Languages of Programs. [*S.l.*: *s.n.*], 2017. p. 1–27.

SINHA, Ravi; MIHALCEA, Rada. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. *In:* IEEE. SEMANTIC Computing, 2007. ICSC 2007. International Conference on. [*S.l.*: *s.n.*], 2007. p. 363–369.

SNOW, Rion; JURAFSKY, Daniel; NG, Andrew Y. Semantic taxonomy induction from heterogenous evidence. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS.

PROCEEDINGS of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. [*S.l.*: *s.n.*], 2006. p. 801–808.

SORATO, Danielly; FILETO, Renato. Linguistic Pattern Mining for Data Analysis in Microblog Texts using Word Embeddings. *In:* XV Brazilian Symp. on Information Systems (SBSI). Aracaju, Sergipe, Brazil: [*s.n.*], 2019.

SORATO, Danielly; GOULARTE, Fábio B *et al.* Análise de Métodos e Ferramentas para Reconhecimento de Palavras Relevantes em Microblogs. *In:* PROCEEDINGS of the XII Brazilian Symposium on Information Systems. [*S.l.*: *s.n.*], 2016.

SPEAKS, Jeff. Theories of Meaning. *In:* ZALTA, Edward N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Spring 2016. [*S.l.*]: Metaphysics Research Lab, Stanford University, 2016.

STRAPPARAVA, Carlo; GLIOZZO, Alfio; GIULIANO, Claudiu. Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3. *In:* PROCEEDINGS of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. [*S.l.*: *s.n.*], 2004.

SUNDERMANN, C. V. *et al.* Exploration of Word Embedding Model to Improve Context-Aware Recommender Systems. *In:* INTERNATIONAL Conference on Web Intelligence (WI). [*S.l.*: *s.n.*], 2018. Disponível em: `https://ieeexplore.ieee.org/abstract/document/8609619`.

SUVITHA, D Sindhu; JANARTHANAN, R. Enriched semantic information processing using WordNet based on semantic relation network. *In:* IEEE. COMPUTING, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on. [*S.l.*: *s.n.*], 2012. p. 846–851.

TAKAMURA, Hiroya; YOKONO, Hikaru; OKUMURA, Manabu. Summarizing a Document Stream. *In:* SPRINGER. ECIR. [*S.l.*: *s.n.*], 2011. p. 177–188.

TJONG KIM SANG, Erik F; BUCHHOLZ, Sabine. Introduction to the CoNLL-2000 shared task: Chunking. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7. [*S.l.*: *s.n.*], 2000. p. 127–132.

TJONG KIM SANG, Erik F; DE MEULDER, Fien. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. [*S.l.*: *s.n.*], 2003. p. 142–147.

TORAL, Antonio; MONACHINI, Monica. Named entity wordnet. *In:* CITESEER. IN Proceedings of the 6th International Conference on Language Resources and Evaluation. [*S.l.*: *s.n.*], 2008.

TRAVAGLIA, Luiz Carlos. **Gramática e interação: uma proposta para o ensino de gramática. rev**. [*S.l.*]: São Paulo: Cortez, 2003.

TURNEY, Peter D. Word sense disambiguation by web mining for word co-occurrence probabilities. **arXiv preprint cs/0407065**, 2004.

TURNEY, Peter D; PANTEL, Patrick. From frequency to meaning: Vector space models of semantics. **Journal of artificial intelligence research**, v. 37, p. 141–188, 2010.

TWITTER. **Quarterly results: Letter to Shareholders**. [*S.l.*], 2018. Disponível em: `https://investor.twitterinc.com/home/default.aspx`.

USBECK, Ricardo *et al.* AGDISTIS-graph-based disambiguation of named entities using linked data. *In:* SPRINGER. INTERNATIONAL Semantic Web Conference. [*S.l.*: *s.n.*], 2014. p. 457–471.

VÉRONIS, Jean. Hyperlex: lexical cartography for information retrieval. **Computer Speech & Language**, Elsevier, v. 18, n. 3, p. 223–252, 2004.

VIAL, Loıc; LECOUTEUX, Benjamin; SCHWAB, Didier. Sense Embeddings in Knowledge-Based Word Sense Disambiguation. *In:* 12TH International Conference on Computational Semantics. [*S.l.*: *s.n.*], 2017.

VOSECKY, Jan *et al.* Integrating social and auxiliary semantics for multifaceted topic modeling in Twitter. **ACM Transactions on Internet Technology (TOIT)**, ACM, v. 14, n. 4, p. 27, 2014.

VOUTILAINEN, Atro. NPtool, a detector of English noun phrases. **arXiv preprint cmp-lg/9502010**, 1995.

VU, Xuan-Son *et al.* Lexical-semantic resources: yet powerful resources for automatic personality classification. **arXiv preprint arXiv:1711.09824**, 2017.

WALLACH, Hanna M. Conditional random fields: An introduction. **Technical Reports (CIS)**, p. 22, 2004.

WANG, Wei; STEWART, Kathleen. Spatiotemporal and semantic information extraction from web news reports about natural hazards. **Computers, Environment and Urban Systems**, Elsevier, v. 50, p. 30–40, 2015.

WEISS, Sholom M *et al.* **Text mining: predictive methods for analyzing unstructured information**. [*S.l.*]: Springer Science & Business Media, 2010.

WU, Zhuofeng *et al.* Identify Shifts of Word Semantics through Bayesian Surprise. *In:* ACM. THE 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. [*S.l.*: *s.n.*], 2018. p. 825–834.

XIA, Chaolun *et al.* What Is New in Our City? A Framework for Event Extraction Using Social Media Posts. *In:* SPRINGER. PACIFIC-ASIA Conference on Knowledge Discovery and Data Mining. [*S.l.*: *s.n.*], 2015. p. 16–32.

XU, Wei *et al.* A preliminary study of tweet summarization using information extraction. **NAACL 2013**, p. 20, 2013.

ZELENKO, Dmitry; AONE, Chinatsu; RICHARDELLA, Anthony. Kernel methods for relation extraction. **Journal of machine learning research**, v. 3, Feb, p. 1083–1106, 2003.

ZHAI, ChengXiang; MASSUNG, Sean. **Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining**. [*S.l.*]: Morgan & Claypool, 2016.

ZHANG, Wei *et al.* Entity Linking Leveraging: Automatically Generated Annotation. *In:* PROCEEDINGS of the 23rd International Conference on Computational Linguistics. Beijing, China: Association for Computational Linguistics, 2010. (COLING '10), p. 1290–1298. Disponível em: `http://dl.acm.org/citation.cfm?id=1873781.1873926`.

ZHONG, Zhi; NG, Hwee Tou. It makes sense: A wide-coverage word sense disambiguation system for free text. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the ACL 2010 System Demonstrations. [*S.l.*: *s.n.*], 2010. p. 78–83.

ZOCK, Michael. Natural languages are flexible tools; that?s what makes them hard to explain, to learn and to use. **Zock and Sabah [1215]**, p. 181–196, 1988.