Uma metodologia de ponderação de sinônimos para geração automática de exercícios de vocabulário

Wander Inácio de Souza
Universidade Federal de Ouro Preto

Orientador: Álvaro Rodrigues Pereira Júnior

UNIVERSIDADE FEDERAL DE OURO PRETO

Uma metodologia de ponderação de sinônimos para geração automática de exercícios de vocabulário

Wander Inácio de Souza
Universidade Federal de Ouro Preto

Orientador: Álvaro Rodrigues Pereira Júnior

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto para obtenção do título de Mestre em Ciência da Computação

Ouro Preto, janeiro de 2017

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S729u Souza, Wander Inácio de.
Uma metodologia de ponderação de sinônimos para geração automática de exercícios de vocabulário. [manuscrito] / Wander Inácio de Souza. - 2017.
69 f.: il.: , gráf., tab..

Orientador: Prof. Dr.´Álvaro Rodrigues Pereira Júnior.
Dissertação (Mestrado Acadêmico). Universidade Federal de Ouro Preto. Departamento de Computação. Programa de Pós-Graduação em Ciência da Computação.
Área de Concentração: Ciência da Computação.

1. Processamento de Linguagem Natural. 2. Análise textual. 3. Geração de exercícios. I. Júnior,´Álvaro Rodrigues Pereira. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004

Bibliotecário(a) Responsável: Luciana De Oliveira - SIAPE: 1.937.800

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO
PROGRAMA DE POS-GRADUACAO EM CIENCIA DA COMPUTACAO

FOLHA DE APROVAÇÃO

Wander Inácio de Souza

Uma metodologia de ponderação de sinônimos para geração automática de exercícios de vocabulário

Dissertação apresentada ao Programa de Pós Graduação em Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Mestre em Ciência da Computação

Aprovada em 30 de janeiro de 2017

Membros da banca

Prof. Dr. Álvaro Rodrigues Pereira Júnior - Orientador Universidade Federal de Ouro Preto
Prof. Dr. Luiz Henrique Campos Merschmann - Universidade Federal de Lavras
Profa. Dra. Lucelene Lopes - Pontifícia Universidade Católica Rio Grande do Sul

Prof. Dr. Álvaro Rodrigues Pereira Júnior, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 15/10/2019

Documento assinado eletronicamente por **Puca Huachi Vaz Penna**, **COORDENADOR(A) DE CURSO DE PÓS-GRADUACÃO EM CIÊNCIA DA COMPUTAÇÃO**, em 08/03/2022, às 14:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0288981** e o código CRC **15D64C1B**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: (31)3559-1641 - www.ufop.br

Dedico este trabalho a Deus, aos meus pais Waldenês *e* Maria Creuza*, aos meu irmãos* Bruno *e* Hugo *e aos meus amigos.*

Uma metodologia de ponderação de sinônimos para geração automática de exercícios de vocabulário

Resumo

Aprender um novo idioma é algo fundamental no mundo globalizado de hoje. O inglês se destaca como o idioma mais estudado atualmente por ser mais utilizado na produção das mais diversas mídias, nas quais cita-se filmes, músicas, jogos, seriados, entre outros. A necessidade de aprender um novo idioma e, principalmente, a língua inglesa, vem impulsionando a criação de novos métodos de aprendizados. Busca-se principalmente uma maior comodidade ao estudante, como a possibilidade de estudar em casa, através da *Internet*. Entretanto, as metodologias de ensino continuam padronizadas e engessadas, não considerando a individualidade de cada aluno no processo de aprendizagem. Uma metodologia de ensino que seja adequada a individualidade de cada aluno demanda de métodos capazes de gerar conhecimento ao estudante por meio do uso de temas que sejam de seu interesse. Dessa forma, neste trabalho propõe-se o desenvolvimento de uma metodologia de construção de ponderação de sinônimos para geração automática de exercícios de vocabulário, no intuito de automatizar a geração de exercícios de vocabulário. O estudo mostrou a viabilidade da aplicação da ponderação automática dos sinônimos e da escolha adequada das palavras utilizadas para o problema de geração de exercícios da língua inglesa. Em especial, para o melhor cenário estudado, a acurácia de seleção alcançou 84,8%, o que parece ser um resultado satisfatório para que se possa, no futuro, gerar uma aplicação real para a geração automática de exercícios de vocabulário.

A methodology for automatic generation of vocabulary exercises

Abstract

Learning a new language is fundamental in the globalized world we live in. English stands out as the most studied language nowadays, mostly because of its use in production of the most diverse media, like films, music, games, series, among others. The need to learn a new language, and especially the English language, has been driving the development of new learning methods. However, teaching methodologies remain standardized and embedded, not considering the individuality of each student in the learning process. A teaching methodology that is adequate to the individuality of each student demands methods capable of generating knowledge to the student through the use of topics that are of interest to them. In this way, this work proposes the development of a methodology for automatic generation of vocabulary exercises, in order to enable the generation of exercises from a document used as input. The study showed the feasibility of the application in automatic generation, reaching the precision of 100% for given scenarios.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Mesmo não sendo a língua mais falada do mundo, o inglês é o idioma de maior importância no cenário mundial. Ele se destaca como opção de segunda língua, devido ao grande volume de conteúdo criado e disponibilizado, independente do meio, para o público que a utiliza. Esse fato é destacado ao observarmos as produções, tais como: jogos, filmes, livros, seriados, programas de computadores, entre outros, que utilizam o inglês como idioma principal.

A presença do inglês no dia a dia sugerem uma maior necessidade de que mais pessoas busquem conhecer esse idioma e aumentar o conhecimento nele. Indivíduos que querem aprender uma língua estrangeira usualmente procuram por uma escola de idiomas. Escolas normalmente ensinam seus alunos utilizando um método crescente no qual, inicialmente, apresentam os termos base para criação do vocabulário. À medida que os estudos avançam, novos termos e formas gramaticais são incorporados (Ramos; 2014). Métodos tradicionais de ensino são os mais utilizados. Entretanto, as diferenças entre os indivíduos, em conjunto com a necessidade de adaptação de algumas técnicas, têm impulsionado o desenvolvimento de novos métodos de ensino (Conte et al.; 2012; Ramos; 2014).

Para pessoas que não estudam uma nova língua de maneira tradicional, o contato com uma nova língua se dá por meio de termos incorporados ao português, devido ao seu uso no cotidiano. Outro ponto de intercessão ocorre quando os indivíduos buscam consumir algum conteúdo, como filmes ou séries, no novo idioma. O uso de determinado conteúdo com frequência denota uma facilidade de o usuário identificar termos frequentes a esse tema.

De modo a alterar esse padrão, apresenta-se a necessidade de um sistema que auxilie o usuário no processo de aprendizagem de um novo idioma, mas que utilize temas de interesse do usuário nesse processo. Partindo desse conceito, é sugerido uma metodologia capaz de utilizar textos informados pelo usuário no processo de ensino de um novo idioma. Com isso, espera-se desenvolver uma metodologia que se difere das tradicionais que não realizam a atividade de personalização do conteúdo utilizado no processo de ensino.

Este trabalho utiliza a hipótese de que cada um dos sinônimos de uma palavra tem significados de maior ou menor relevância, de acordo com o contexto no qual é utilizado. Logo, é possível ponderar os sinônimos de uma palavra para cada sentido. A ponderação define que, por exemplo, "*cube*" quando utilizado no contexto "*a solid bounded by six equal squares, the angle between any two adjacent faces being a right angle*" tem uma maior relação de sinonímia com "*square*" do que "*multiply*". Em contrapartida, quando "*cube*" está sob o contexto de "*the third power of a quantity, expressed as* $a^3 = a{\cdot}a{\cdot}a$", a palavra "*multiply*" torna-se aquela com sinonímia mais forte.

A ponderação da lista de sinônimos de uma palavra, a partir das suas definições, permite identificar relações entre os sentidos e os sinônimos. A partir dessa definição, torna-se possível, também, utilizar da relação de sinonímia entre os termos para a geração de exercícios automatizados de vocabulário.

1.1 Objetivos

O objetivo deste trabalho consiste em desenvolver uma metodologia para ponderação de sinônimos de diferentes definições de um mesmo termo, visando a geração automática de exercícios de vocabulário de uma língua.

Os objetivos específicos são:

- Coleta de bases de dados de dicionários e *thesaurus* da língua inglesa;
- Derivar uma base de conhecimento de vocabulário, definindo o nível de sinonímia de dois termos por meio da ponderação de suas definições;
- Desenvolver um modelo de associação de conteúdo de acordo com os diferentes níveis de conhecimento do vocabulário da língua que cada usuário possui;

- Desenvolver um método de geração de exercícios de vocabulário, utilizando os dados presentes na base de conhecimento;
- Validar os resultados obtidos pelo método de ponderação de sinonímia;
- Validar os resultados obtidos pelo método de geração de exercícios.

1.2 Justificativa

A predição do conhecimento de um indivíduo sobre o vocabulário de um idioma permite alocar a pessoa na posição em que deve iniciar, ou continuar, seus estudos. Os métodos preditivos atuais pertencem às escolas de idiomas, utilizados com alunos que seguem sua metodologia e querem estudar nessas escolas. Caracterizar o indivíduo de acordo com seus interesses, juntamente com o seu nível de conhecimento, permite um direcionamento de conteúdo mais específico. É desconhecida alguma metodologia que realize esse tipo de atividade.

A necessidade de compreender o nível de conhecimento em língua inglesa do indivíduo é um dos fatores que justificam a realização deste trabalho. A partir da caracterização, torna-se possível um nivelamento do estudante e, com isso, auxiliá-lo a evoluir seus conhecimentos em uma nova língua. O inglês foi selecionado como idioma a ser aprendido devido à sua importância no cenário mundial. O destaque da língua em relação à produção de conteúdos, dos mais diversos interesses, conforme supracitado, também identifica a necessidade de aprender este idioma.

Outro fator que justifique a realização deste trabalho consiste na possibilidade de variação do conteúdo apresentado a cada estudante. Com esse dinamismo no processo de ensino, torna-se possível a utilização de planos de ensino personalizados por aluno. Assim, o processo de aprendizado utiliza de temas de interesse do estudante para auxiliar no seu desenvolvimento.

Essa motivação é relevante porque, atualmente, os métodos tradicionais de aprendizado de segunda língua estão diretamente ligados às escolas de idiomas. Mesmo sendo o meio mais comum de estudo, existem alguns fatos que o tornam desinteressante: o conteúdo estudado é fechado e pouco diversificado, os cursos têm alto custo financeiro, grande período de duração, além da necessidade de deslocamento para a escola. Essas características podem ser fatores desmotivantes para o aluno.

A partir das evoluções nas metodologias de ensino, suprimiu-se a necessidade de deslocamento para a escola de idiomas por meio do surgimento de ferramentas como o Duolingo[1], *My English Online*[2], *Open English*[3], *English Town*[4], entre outras. Essas ferramentas permitem estudar em casa e, em alguns casos, com o auxílio de professores. Mesmo com essa flexibilidade parcial, ainda observa-se que os conteúdos são fixos, ou seja, todos os alunos seguem o mesmo conteúdo para aprendizado. Mesmo com a criação de novos métodos de ensino, ainda são desconhecidos aqueles que utilizem dados públicos como parte do aprendizado, seja para ensinar ou para exemplificar. Técnicas que utilizam do interesse específico do usuário para auxiliar no aprendizado também são desconhecidas.

Acredita-se que o uso de diferentes *corpora*, definidos pelos usuários e pertencentes a diversas áreas de interesse, possibilitarão um maior volume de conteúdos a serem estudados. Juntamente a esse fato, existe uma simpatia do usuário com o tema estudado, o que espera-se que aumente o interesse do mesmo pelo estudo e, consequentemente, uma melhora no processo de estudo.

1.3 Organização da Dissertação

Os próximos capítulos estão divididos da seguinte maneira: no Capítulo 2, é detalhada a fundamentação teórica. No capítulo 3, são apresentados os trabalhos relacionados. Em seguida, no capítulo 4, é apresentado o processo de manipulação das bases de dados para caracterizar o nível de conhecimento do usuário. No capítulo 5, é proposta a metodologia de extração de conhecimento e geração de exercícios. Em seguida, o capítulo 6 apresenta os resultados experimentais obtidos pela metodologia de geração de exercícios. Finalmente, o capítulo 7 apresenta as conclusões obtidas neste trabalho e os trabalhos futuros sugeridos.

[1] https://pt.duolingo.com/
[2] http://www.myenglishonline.com.br/
[3] http://ww.openenglish.com.br
[4] http://ww.englishtown.com.br

Capítulo 2

Fundamentação teórica

Este capítulo apresenta a fundamentação teórica, quando são descritos conceitos importantes associados ao conteúdo deste trabalho. Inicialmente, na Seção 2.1, explicitam-se conceitos referentes à Recuperação da Informação. Em seguida, a Seção 2.2 detalha os conceitos associados ao processamento de linguagem natural. Por fim, na Seção 2.3, são expostos conceitos referentes à Desambiguação de Sentido dos termos.

2.1 Recuperação de Informação

Dentro da área da Ciência da Computação, a Recuperação de Informação (RI) é responsável por trabalhar com o armazenamento e identificação de documentos, além de possibilitar a recuperação de informação contida neles. Espera-se que um Sistema de RI seja capaz de trabalhar com bases de dados de diversos tamanhos, possibilitando que essas atividades sejam realizadas de forma automática (Baeza-Yates and Ribeiro-Neto; 1999; Salton and McGill; 1983).

Os documentos utilizados pela RI geralmente não contam com uma estruturação bem definida. Isso indica que: (1) os dados estão armazenados em texto puro ou; (2) estão organizados de forma semiestruturada com, por exemplo, divisão em *tags* ou blocos (Chowdhury; 2010; Manning et al.; 2008). Dentre as tarefas realizadas no processo de RI, destacam-se a coleta, a indexação e a consulta.

2.1.1 Palavras, Termos e Tokens

Dentro do contexto destacado na Seção 2.1, palavra, termo e *token* têm definições diferentes, cada uma sendo citada a seguir.

Manning et al. (2008) define uma palavra como uma concatenação limitada de caracteres, que tenha sentido de uso em uma língua, independente da grafia utilizada. Uma palavra pode conter caracteres em caixa alta ou baixa, acentuação ou representar qualquer tempo verbal, por exemplo. Cada palavra é única para a coleção, independente do número de vezes que é utilizada.

Termo é a versão normalizada da palavra. Essa normalização é definida pelo sistema em que o modelo foi implantando. Alguns exemplos de normalização são: retirar acentuações e utilizar apenas caracteres de caixa baixa. Cada termo é único para a coleção, independente do número de vezes que é utilizado (Manning et al.; 2008).

O *token* representa cada instância, ou uso, das palavras, em um documento. Por exemplo, cada vez que a palavra "love" é utilizada em um documento, é contabilizado um novo *token* dessa palavra. Cada termo identificado em um documento contém pelo menos um *token* (Manning et al.; 2008).

2.1.2 Coleta

A coleta é a tarefa responsável pela geração de uma base local dos dados encontrados na *internet* (Manning et al.; 2008). A realização da coleta é explicada através de duas etapas.

A primeira consiste na identificação do conteúdo de interesse. Para isso é utilizado um *WebCrawler* que consiste em um sistema capaz de visitar as páginas na *internet*. A navegação é sequencial, ou seja, inicialmente são definidos pontos de partida (ou sementes). Cada semente representa uma página que pode ser visitada e contém o endereço, ou *hyperlink*, para outras páginas que podem ser acessadas a partir da atual (Manning et al.; 2008).

Em cada página visitada é realizada a segunda tarefa, coleta. Nela, deve-se identificar se o conteúdo faz parte do contexto de RI trabalhado e, caso positivo, a página é armazenada localmente. Alguns pré-processamentos devem ser realizados durante a coleta de modo a: (1) identificar e descartar páginas com conteúdo duplicado; (2) identificar e

descartar páginas que sejam *spams* (Manning et al.; 2008).

Essa é a tarefa inicial para a realização do processo de RI na *internet*, pois, a partir dela, é gerada a base de dados local utilizada nos processos subsequentes. Existem casos nos quais é necessário realizar a coleta periodicamente para manter a base de dados local atualizada. Uma vez formada a base de dados inicial, todas as demais tarefas de RI devem ser executadas (Chowdhury; 2010; Manning et al.; 2008).

2.1.3 Indexação

A indexação é a tarefa executada após a coleta. Ela é responsável por extrair termos significativos, que representem o documento, e armazená-los de modo a facilitar a recuperação. A execução é dividida em três etapas: (1) *tokenização* do texto; (2) processamento linguístico; (3) indexação (Manning et al.; 2008).

A *tokenização* consiste em transformar o documento em uma lista de *tokens*, ou palavras, que sejam capazes de identificar o documento. Comumente, nesse processo, são descartados os termos mais comuns da língua, as *stopwords*, uma vez que não auxiliam na identificação do documento (Baeza-Yates and Ribeiro-Neto; 1999; Chowdhury; 2010). Não existe uma lista predefinida de *stopwords*, neste caso são definidos conjuntos diferentes para aplicações diferentes. Por serem palavras que não auxiliam no processamento da coleção, elas têm uma frequência muito grande na coleção. Por exemplo, artigos estão presentes em todos os documentos e normalmente podem ser pertencer a este conjunto. Outro é exemplo consiste em uma coleção de documentos de computação. Neste casos, a palavra computador pode ser considerada uma *stopword*, uma vez que deve estar presente em uma grande porção de documentos.

Com essa lista gerada, é realizado um processamento linguístico, a fim de normalizar os termos. Este processo busca minimizar os possíveis ruídos durante o processamento de linguagem natural. Dentre os processos que podem ser executados, destacam-se a lematização (redução do termo à sua forma básica de dicionário, o lema) e o *stemming* (busca identificar o radical dos termos). Além dessas, podem ser executados outros processamentos como: remoção de caracteres acentuados, remoção de plurais ou outros processamentos que auxiliem na recuperação. Em alguns casos, pode não ser necessário realizar pré-processamentos na coleção. A necessidade do processamento linguístico é analisada para cada sistema de RI (Manning et al.; 2008).

Ao final das etapas anteriores, é realizada a geração do índice. A estrutura mais

comum para seu armazenamento é a lista invertida. Nela, cada termo é alocado seguido da relação de documentos nos quais o termo está presente. Essa lista pode conter outras informações pertinentes ao contexto no qual será utilizado como: frequência do termo dentro do documento, enumeração das posições em que o termo está presente no documento, entre outros (Baeza-Yates and Ribeiro-Neto; 1999; Chowdhury; 2010; Manning et al.; 2008).

2.1.4 Consulta

Na fase da consulta, o usuário realiza a busca nos dados coletados e indexados, a fim de obter a informação ali armazenada. A consulta é dividida em três processos distintos. No primeiro, o usuário explicita o que deseja encontrar, normalmente por meio de termos-chave. Em seguida, o sistema de RI busca identificar quais os documentos que armazenam as informações que o usuário deseja. Essa validação ocorre calculando a similaridade entre os termos-chave, especificados na consulta, e os termos indexados de cada um dos documentos da base. No último processo são retornados os documentos identificados como relevantes para a consulta especificada (Baeza-Yates and Ribeiro-Neto; 1999; Manning et al.; 2008).

O resultado pode ser apresentado de vários modos. Comumente, é gerada uma lista ordenada com os documentos de maior relevância. O *ranking* é gerado a partir da ordenação dos documentos recuperados de acordo com a relevância para a consulta e o intuito é auxiliar o usuário a identificar quais os documentos podem ser mais relevantes para a consulta (Chowdhury; 2010; Manning et al.; 2008).

2.1.5 Coleções de Documentos (corpora)

Os Sistemas de RI realizam suas atividades em coleções de documentos que são agrupamentos de documentos independente dos temas e da estruturação presente de cada um. Elas podem ser armazenadas em repositórios locais ou externos. Dentre os tipos de coleções de documentos, destaca-se o *corpus*, utilizado neste trabalho.

Na Linguística, um *corpus* consiste em uma base de linguagem natural, geralmente utilizado para pesquisa linguística. A formação dele é realizada a partir de dados autênticos, utilizando critérios linguísticos específicos, mas abrangente o suficiente para ser representativo. Sua armazenagem deve existir em formato eletrônico e sua criação

segue, comumente, temas específicos (Sardinha; 2004). Um conjunto de diversos *corpus* é chamado *corpora*.

A representação dos dados de um *corpus*, ou coleção de documentos, é chamada vocabulário. Em uma descrição mais simples, o vocabulário é a lista de todos os termos presentes na coleção de documentos (Manning et al.; 2008).

Dicionário consiste na estrutura de dados responsável por armazenar o vocabulário. Nele, é armazenada cada uma das palavras do vocabulário, juntamente com as demais informações relevantes para o sistema de RI. A estrutura é composta por uma lista de termos. Cada termo contém uma lista de documentos a qual ele está presente (chamada *postings*). Podem ser incluídas informações adicionais a cada documento, como: a frequência do termo, suas posições, entre outras (Manning et al.; 2008).

2.2 Processamento de Linguagem Natural

Linguagem natural é a forma de comunicação do ser humano. A estrutura é complexa, pois ela pode variar de acordo com a proposta na qual está sendo utilizada. Por exemplo, a linguagem natural pode ser utilizada para informar, corrigir, direcionar ou solicitar a realização de uma ação, direcionar reações físicas ou cognitivas, entre diversas funções (Baeza-Yates et al.; 2015). O Processamento de Linguagem Natural (PLN) busca compreender computacionalmente todo o conceito atribuído a frases de linguagem natural.

De modo a compreender esse processo, inicialmente, a Seção 2.2.1 apresenta os conceitos fundamentais do Processamento de Linguagem Natural. Em seguida, a Seção 2.2.2 demonstra como dicionários e definições de dicionários são aplicados neste processo. A Seção 2.2.3 conceitualiza sobre sinonímia. Por fim, a Seção 2.2.4 apresenta a base de dados *WordNet* e a sua importância para a evolução do Processamento de Linguagem Natural.

2.2.1 Conceitos Fundamentais de Processamento de Linguagem Natural

Conceitualmente, o processamento automático de linguagem consiste no uso de um computador para processar qualquer tipo de linguagem (Bobrow et al.; 1967). Entretanto,

diversos autores delimitam o escopo da área para linguagem natural, eliminando linguagens artificiais como linguagens de programação (Bobrow et al.; 1967).

A partir desses conceitos básicos, houve novos estudos sobre o PLN. Damerau (1976) é o primeiro autor a destacar que simulações linguísticas são mais realizadas pelos departamentos de Ciência da Computação e Psicologia ao invés da Linguística. Nesse ponto, é observada uma divisão no estudo da Linguística, na qual a organização e conceitos são responsabilidade da Linguística Teórica. Complementarmente, a Linguística Computacional utiliza algoritmos sobre coleções de documentos de linguagem. Por fim, o Processamento de Linguagem Natural estuda o uso do computador, a fim de extrair ou reorganizar informações em coleções de dados (Damerau; 1976).

Uma definição mais atual caracteriza o PLN como o uso de computadores em uma base de dados de linguagem natural, visando o seu processamento e manipulação, a fim de obter dados úteis (Chowdhury; 2003). A evolução do PLN demanda a existência de bases em diversas áreas de pesquisa, como: Ciência da Computação, Ciência da Informação, Matemática, Engenharias, entre outras. Assim, é possível realizar processamentos direcionados a determinada área, visando uma melhor qualidade resultante do processo.

2.2.2 Dicionários e Definições de Dicionário

Na Linguística, o modo de catalogar os termos estudados é por meio de dicionários. Um dicionário consiste em uma compilação total ou parcial de unidades léxicas organizadas em uma ordem convencionada (Houaiss; 2003). Sendo total ou parcial, a definição também se estende ao conjunto de vocábulos agrupados por um indivíduo, em uma época, ou ainda, termos e referências sobre determinado tema. Sendo representado por um conjunto de termos, o que diferencia um dicionário de glossários ou vocabulários, que também representam compilações de termos, é o conjunto de definições associadas ao termo do dicionário.

O termo definição tem como significado: definição exata ou um estabelecimento de limites (Houaiss; 2003). A associação das definições a um conjunto de termos forma um dicionário. Com isso, constata-se que definição delimita o conjunto de significados que um termo tem associado a si.

É sabido que a linguagem natural consiste em um conjunto de termos, formando frases, que contém um significado associado. Conforme supracitado, cada termo presente na

frase contém um conjunto de significados agregados. A área de pesquisa responsável por identificar qual a definição associada ao termo dentro de determinada frase é chamada *Word Sense Disambiguation*, ou WSD (Desambiguação Lexical de Sentido).

2.2.3 Sinonímia

Dentro da Linguística, existem palavras que têm definições aproximadas. Essa relação é chamada de sinonímia e o Dicionário Houaiss (2003) a define como a relação de sentidos de dois termos com significados próximos.

A relação de sinonímia é amplamente difundida, na qual um termo contém uma lista de sinônimos. Entretanto, é desconhecido um estudo que determina essa relação diretamente a partir das definições dos termos relacionados. Ou seja, existe a relação de sinonímia entre dois termos, mas não há associação entre as definições dos termos que determinam essa associação.

2.2.4 A Base de Dados de Referência WordNet

A *WordNet*[1] é uma base de dados léxica amplamente utilizada em trabalhos relacionados ao PLN. Isso ocorre, principalmente, por ser gratuita e disponibilizar uma estrutura relevante para pesquisas de Linguística Computacional e Processamento de Linguagem Natural. Ela provê uma rede de palavras significativamente correlacionadas na qual substantivos, adjetivos, verbos e advérbios são agrupados em conjuntos de sinônimos cognitivos. Com isso, as palavras são agrupadas de acordo com seus significados, nos chamados *synsets*. Não há a presença de qualquer outro padrão na base de dados, logo, os *synsets* formam a estrutura básica do *WordNet*.

Além da relação de sinonímia entre os termos, há a correlação entre os grupos formados. Essa correlação entre os grupos de sinônimos é chamada *hyponymy* e busca relacionar palavras que não são sinônimas, mas existe uma relação entre elas. Por exemplo, no *synset* de "móveis" não está presente o termo "cama", entretanto, há uma relação entre essas palavras, uma vez que uma cama é um móvel. Logo, é criada uma relação entre os grupos de sinônimos de móveis e cama, o chamado *hyponymy*. As relações de sinonímia existentes entre os diversos termos, é um facilitador para a metodologia proposta. Isso ocorre porque os *synsets* apresentam a lista de sinônimos para determi-

[1] https://wordnet.princeton.edu/

nado termo, entretanto não existe uma ponderação entre os sinônimos, o que torna este trabalho relevante.

Outro fator relevante que sugere o uso do *WordNet* como a base de dados de referência é a sua API (*Application Programmin Interface*), que oferece diversos serviços relacionados ao uso de dicionários, como: busca de termos, definições, antônimos, sinônimos, exemplos de uso de um termo em suas definições, além da relação entre todos esses componentes. A partir dela, há a facilidade de identificar os diversos dados relacionados ao termo em uma única base e sem a necessidade de processamentos adicionais.

Uma vez definida a base de dados de entrada, é iniciado o detalhamento do Extrator de Conhecimento. Para isso, detalha-se inicialmente o Extrator de Dados do *WordNet*.

2.3 Desambiguação Lexical de Sentido

A Desambiguação Lexical de Sentido (DLS) ou *Word Sense Desambiguation* (WSD) consiste em identificar o sentido adequado de uma palavra em um dado contexto. O conjunto de sentidos de um termo é um conjunto finito. As pesquisas se iniciaram no contexto de DLS com o intuito de melhorar a automação de tarefas de tradução automatizada por computadores (Agirre and Edmonds; 2007).

Estudos relacionados à DLS têm sido realizados continuamente e novas soluções têm sido propostas avançando o estado da arte. Isso ocorre devido ao dinamismo envolvido nessa área, principalmente referente à evolução do vocabulário, novas palavras são constantemente inseridas nos dicionários e consequentemente novas definições são associadas a esses termos.

O primeiro estudo sobre a automatização da DLS foi realizado por Lesk (1986). Ele utilizou de dicionários e foi proposto para ser independente de idiomas, ou seja, a partir de um conjunto de dicionários, seu método, originalmente proposto para a língua inglesa, pode ser utilizado para desambiguação lexical em qualquer idioma.

Conforme supracitado, as pesquisas em DLS têm evoluído e novas técnicas e recursos têm sido aplicados de modo a melhorar a desambiguação lexical. Atualmente, as tarefas realizadas em sistemas DLS variam por domínio, entretanto, elas podem ser modificadas ou agrupadas para serem utilizadas em casos específicos (Agirre and Edmonds; 2007; Plaza and Diaz; 2011). São três as principais tarefas realizadas (Martin and Jurafsky; 2000), a saber:

- *Lexical Sample*: a DLS é responsável por desambiguar um conjunto previamente determinado de palavras, ou conjunto amostra, especificando os seus possíveis sentidos;
- *All words*: todas as palavras presentes em um texto devem ser desambiguadas;
- Desambiguação por transferência: a DLS deve desambiguar uma palavra e, em seguida, determinar a melhor tradução dessa palavra em outro idioma.

Além das tarefas, outros conceitos são aplicados no processo de desambiguação. Esses conceitos são características utilizadas pela DLS, a fim de obter um melhor resultado ao final da desambiguação (Agirre and Edmonds; 2007; Martin and Jurafsky; 2000; Plaza and Diaz; 2011). Os principais citados na literatura relacionada são:

- *Bag of words*: é o conjunto de palavras que circundam a palavra analisada. Nesse método de avaliação, há uma variação do número de palavras a serem analisadas;
- Classe gramatical: representa a classe gramatical da palavra analisada, um dado importante, uma vez que palavras apresentam definições diferentes quando utilizadas com classes gramaticais diferentes;
- Similaridade semântica: valor numérico que indica o quanto o sentido de duas palavras é próximo. A similaridade semântica não necessariamente indica que os termos são sinônimos, mas sim o quanto os dois termos se aproximam semanticamente.
- Contexto: representa o cenário no qual a palavra está sendo utilizada. Normalmente é associado às palavras em volta da palavra analisada.

Sendo algoritmos de alta complexidade computacional, normalmente as tarefas de desambiguação de sentidos estão associadas a heurísticas. Estas buscam realizar a tarefa de desambiguação de sentido a um custo computacional menor do que algoritmos não heurísticos. As heurísticas tendem a reutilizar um sentido já previamente identificado em outras partes do texto. Essa reutilização pode ocorrer por sentença ou parágrafo. Outra heurística consiste em determinar o sentido da palavra de acordo com seu uso. Nesse caso, uma vez identificado o sentido da palavra em um contexto, nas próximas aparições desse contexto, o mesmo sentido será utilizado. Uma terceira heurística consiste na utilização de um mesmo sentido para todas as ocorrências da palavra. Entretanto, nesse

caso, é utilizado um *corpus* anotado como referência para a desambiguação (Agirre and Edmonds; 2007; Mihalcea et al.; 2006). Existem outras heurísticas desenvolvidas para a desambiguação de sentidos, entretanto, todas são variações desses modelos citados.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, serão apresentados trabalhos da literatura que se relacionam ao trabalho proposto neste documento. A Seção 3.1 apresenta trabalhos que utilizam de *corpora* de referência em Processamento de Linguagem Natural . Na Seção 3.2, são explicitadas as técnicas que auxiliam a melhorar o aprendizado dos estudantes. Em seguida, a Seção 3.3 apresenta métodos para a geração automática de exercícios a partir de dados contidos em *corpora*. Por fim, a Seção 3.4 apresenta trabalhos de desambiguação de sentido lexical.

3.1 Uso de Corpus de Processamento de Linguagem Natural

O trabalho de Baeza-Yates et al. (2015) apresenta um algoritmo capaz de predizer a dificuldade de palavras para pessoas com dislexia. Para tanto, foi realizado o treinamento do algoritmo com uma base de dados com termos rotulados de modo binário, fácil e difícil, que identifique a dificuldade da palavra. Foi utilizado um *corpus* de referência criado pelos autores. A partir desse *corpus* se obteve uma acurácia de 72,34%, valor crível na literatura para trabalhos desse âmbito. Partindo do propósito de pesar os termos de acordo com sua dificuldade e visando simplificar textos para pessoas com dislexia, surgiu a necessidade de identificar sinônimos de maneira eficiente. Foi analisada a eficiência de quatro algoritmos: *naive*, que utilizou uma abordagem de força bruta, realizando a comparação entre dois vetores; APATS, que utiliza múltiplos padrões para correspondência de palavras aproximadas; Burkhard-Keller *Tree*, que é uma estrutura de dados baseada em árvore desenhada para encontrar correspondências entre objetos

similares, utilizando espaço métrico; e *Trie* baseada em NDFA, que utiliza representação do dicionário em uma árvore digital para encontrar palavras semelhantes, utilizando expressões regulares. Os testes foram realizados em *corpus* que contenham 10.000, 100.000 e 1.200.000 termos. Os testes demonstraram que o método *Trie* baseado em NDFA é o mais eficiente, obtendo resultados em menos de um segundo.

Rello et al. (2013) estudam duas estratégias para simplificar o conteúdo de textos para pessoas com dislexia. A primeira consiste em substituir a palavra pelo seu sinônimo mais simples e a segunda sugere uma lista de sinônimos. Cada abordagem difere de acordo com a estratégia utilizada: enquanto a primeira ranqueia os sinônimos com base no *OpenThesaurus for Spanish* e seleciona o mais simples; a segunda estratégia, após definir que um determinado termo é complexo para portadores da dislexia, ranqueia os sinônimos de acordo com a simplicidade e exibe os três mais simples, podendo exibir os demais de acordo com as solicitações do usuário. Não foram observadas melhorias significativas em termos de simplificação de compreensão e legibilidade dos textos. A substituição pelo sinônimo mais simples, para pessoas com dislexia, gerou textos mais complexos para pessoas sem o distúrbio. Em contrapartida, o uso da sugestão de lista de sinônimos gerou os melhores resultados para pessoas com dislexia.

O trabalho de Laufer and Waldman (2011) investiga o uso excessivo e incorreto de termos por estudantes de inglês como segunda língua. Ele cita que, a partir das técnicas de análise de *corpus*, foi possível estudar grandes coleções compostas por: textos produzidos pelos estudantes, exemplos de uso da língua por nativos e comparação entre as duas coleções anteriores. Os estudantes foram divididos em três grupos de proficiências diferentes: básico, intermediário e avançado. A partir desse agrupamento, foram estudadas as relações entre cada grupo e os textos escritos por nativos da língua, verificando: a relação entre os números de verbos na coleção de cada um, a relação existente entre a coleção gerada pelos estudantes e a coleção nativa referente ao seu nível de proficiência, e a porcentagem de erros recorrentes para os estudantes. A coleção utilizada foi a IL-CoWE (*Israeli Learner Corpus of Written English*), que foi gerada a partir de textos de estudantes de 9 a 12 anos, juntamente com programas de treinamento de inglês. Os *corpora* são compostos de termos básicos, intermediários e avançados, contendo, respectivamente 41.621, 47.117 e 202.311 palavras. Como resultado da pesquisa, foi verificado que estudantes de segunda língua, de todos os níveis, subutilizam verbos em sua escrita, quando comparados a escritores nativos da mesma idade. Também constatou-se que a escrita de alunos de nível básico é mais conformista com as regras do que dos demais níveis. Esses resultados só foram possíveis devido a análise de *corpora* independentes

para cada nível de estudante analisado.

An et al. (2003) apresentam um método de construir um *corpus* de entidades nomeadas. Para isso é utilizada uma lista de entidades nomeadas (EN) e um sistema de busca web para realizar a coleta de dados. No trabalho, são tratadas apenas as três maiores categorias: pessoas, empresas e locais. A coleta foi realizada apenas em páginas que contenham ao menos uma entidade nomeada; em seguida, a coleção foi processada, a fim de manter apenas as sentenças que continham EN. Foram realizados dois tipos de treinamento para o algoritmo: a partir de uma coleção gerada manualmente e outra gerada automaticamente. A primeira opção gerou os melhores resultados finais. O *corpus* resultante foi obtido a partir das sentenças previamente coletadas, quando a EN de cada frase foi identificada, após o treinamento do algoritmo, e adicionada à coleção.

Jin and Wong (2002) propuseram um método estatístico para identificar padrões de *strings* para a criação de um dicionário em língua chinesa a partir de *corpora* diferentes. A primeira etapa consistiu na montagem do dicionário a partir da extração das palavras, utilizando um método estatísticos incrementado com informações de contexto. A segunda etapa consistiu em indexação do dicionário criado para facilitar na recuperação da informação. Os resultados obtidos foram validados a partir da precisão e revocação para os dois casos: (1) palavras extraídas corretamente; (2) palavras novas extraídas corretamente. A partir disso, obteve-se uma extração de palavras com valores de 96,5% e 94,5% para precisão e revocação, respectivamente; a extração de novas palavras resultou em dados de 85% e 78,1%, respectivamente. Uma análise complementar foi realizada e obteve-se acurácia de 99%. Esses valores indicam resultados satisfatórios para o problema.

Breland (1996) estuda a correlação entre a frequência de termos para quatro *corpus* diferentes e uma lista de palavras ordenadas por dificuldade a partir de testes de vocabulário. Os resultados indicaram alta relação entre a dificuldade da palavra e a sua frequência na coleção, demonstrando que termos mais comuns são de mais fácil entendimento do que termos que são utilizados com menor frequência. O resultado demonstra forte relação entre frequência de uso de um termo com sua dificuldade.

Os trabalhos apresentados nessa seção se relacionam ao trabalho proposto no uso de *corpus* em suas pesquisas. Dentre eles destaca-se o trabalho de Rello et al. (2013) que utiliza *corpus* de sinônimos em PLN. Este trabalho também utiliza de *corpus* de sinônimos. Diferentemente do trabalho de Rello et al. (2013), utiliza a base original, este trabalho manipula o *corpus* de sinônimos. O resultado do trabalho de Breland

(1996), demonstrando que termos mais comuns são de mais fácil entendimento enquanto termos raros são mais difíceis de serem compreendidos, é utilizado como referência para calcular a dificuldade dos exercícios. Nesse caso, quanto mais raro é o conjunto de termos presente, mais difícil se torna o exercício.

3.2 Técnicas de Aprendizado em Processamento de Linguagem Natural

Carr and Mazur-Stewart (1988) investigam o melhoramento na compreensão e retenção dos termos de um vocabulário a partir da seguinte justificativa: o entendimento do termo em um contexto auxilia na compreensão do texto. O estudo baseia-se em testar a eficiência de aumentar o entendimento e retenção de vocabulário não familiar ao leitor, a partir de dissertações sobre temas pouco familiares aos estudantes. Esse foi o primeiro esforço de ensinar um novo vocabulário sem partir de *corpus* pré-definido. O método utiliza uma coleção sobre o tema a ser usado. Para a validação do mesmo, foram criados dois grupos: um do método tradicional e outro do método proposto. Os resultados do trabalho demonstraram uma maior evolução, tanto imediata quanto a longo prazo, no vocabulário e retenção de conhecimento dos alunos que estudaram utilizando o método proposto.

O trabalho de Sildus (2006) busca investigar o aprendizado a partir de vocabulários interativos. O estudo foi realizado com 272 participantes que foram direcionados randomicamente a dois grupos: controle e experimental. A prática utilizou projetos de vídeos nos quais as equipes planejavam, preparavam e gravavam um vídeo conversando sobre temas específicos. O grupo experimental preparou vídeos sobre desfiles de moda enquanto o grupo de controle realizou atividades comuns à sala da aula, como: preenchimento de lacunas, encontro de correspondências e exercícios de múltipla escolha. Os resultados foram atingidos a partir das notas geradas em testes realizados antes e depois da prática. Os resultados obtidos demonstraram evolução no vocabulário de ambos os grupos, com destaque ao grupo experimental que obteve maior ganho ao longo do tempo. O resultado obtido por Sildus (2006) demonstra que o uso de vocabulários interativos pode auxiliar na evolução do vocabulário do indivíduo.

O trabalho de O'Neil and Perez (2013) apresenta a importância de utilizar a tecnologia para ensinar o vocabulário. Para a realização do estudo, foram utilizados 80

professores de inglês como segunda língua para estudantes que iriam realizar o teste de proficiência da língua inglesa. Foram utilizadas listas de palavras, cartões *flash*, dicionários digitais, dicionários da *internet*, programas de autoria dos professores, dados da *internet*, dados de *blogs* e *wikis*. Os alunos informaram a frequência que utilizam cada uma das ferramentas para estudar o novo idioma e os resultados demonstraram que listas de palavras, cartões *flash* e dicionários da *internet* são os meios comuns de estudo. O'Neil and Perez (2013) demonstrou o interesse dos estudantes em aprender um novo idioma a partir de meios digitais, dado utilizado neste trabalho.

Os trabalhos apresentados nessa seção se relacionam ao trabalho proposto no uso da tecnologia como meio de aprendizado. Todos os trabalhos apresentam propostas distintas mas concluem que o uso da tecnologia pode ser um facilitador em algumas etapas do aprendizado. Destaca-se o trabalho de O'Neil and Perez (2013), que demonstra uma tendência no estudo de novos idiomas a partir da *internet*, característica que este trabalho se propõe a utilizar.

3.3 Geração de Exercícios

Lin et al. (2015) apresentam um *framework* genérico para geração a semiautomática de exercícios. Os grandes diferenciais desse trabalho estão na possibilidade de controlar a dificuldade e por ser um sistema genérico, independente de domínio. O trabalho é realizado utilizando dados do *linked data*[1]. As análises são realizadas sobre a predição da dificuldade do exercício gerado pelo *framework* e demonstram a superação de quatro *baselines* em 50%.

Sakaguchi et al. (2013) propõem um método para a geração de exercícios de preenchimento para aprendizes da língua inglesa. O método proposto busca satisfazer confiabilidade e validade dos resultados obtidos. É empregado um modelo discriminativo utilizando padrões de erros de *corpora* gerados por aprendizes. Os resultados demonstraram que 98,3% dos exercícios gerados pelo método são confiáveis.

O trabalho de Sung et al. (2007) sugere um protótipo para a geração de exercícios com o intuito de testar a compreensão de texto para aprendizes da língua inglesa. O trabalho traduz o texto em uma representação de rede semântica e a melhora iterativamente com conhecimentos intrínsecos, como gramática. O resultado é gerado a partir

[1] http://linkeddata.org

da compreensão de texto, necessidades, proficiência e erros comuns do aprendiz.

Esta seção apresenta trabalhos que propõem técnicas para geração de exercícios, tarefa também utilizada no trabalho proposto. Foram observados modelos diferentes de geração de exercício, que, entretanto, não se adéquam a esse trabalho. A inadequação do trabalho de Lin et al. (2015) se a divergência entre as bases de dados. Enquanto este trabalho utiliza de um banco de dados relacional, Lin et al. (2015) apresenta sua proposta para uso em dados do *linked data*. Sakaguchi et al. (2013) e Sung et al. (2007) propõe exercícios de gramática, com conexão a um texto, ao contrário deste trabalho que apresenta exercícios de vocabulário.

3.4 Desambiguação Lexical de Sentido

Lesk (1986) apresenta a primeira abordagem para desambiguação de sentido lexical. A abordagem utiliza a hipótese de que o sentido da palavra pode ser encontrado a partir do contexto no qual a palavra está sendo utilizada. A sua técnica consiste na sobreposição de palavras entre as definições extraídas do dicionário e na rotulagem das palavras no contexto. O significado escolhido consiste naquele com maiores sobreposições. Sendo uma abordagem independente de língua, é uma técnica que pode ser utilizada em vários idiomas. Por esse mesmo motivo, é a base para diversas novas técnicas de desambiguação propostas nos mais diversos idiomas. A técnica original apresenta uma acurácia entre 50% e 70%, obtidos de testes manuais realizados pelo autor. Não sendo um resultado expressivo a partir de comparações com novas hipóteses, o algoritmo é amplamente utilizado como *baseline* para novas técnicas.

O trabalho de Mihalcea and Moldovan (1999) apresenta uma abordagem de desambiguação baseando-se no uso de pares de palavras para a construção de consultas *web*. A hipótese de trabalho é baseada na maior similaridade entre as definições, ou seja, o sentido correto para um termo consiste naquele com maior similaridade com o contexto no qual está sendo utilizado. O *corpus* utilizado pelos autores é a própria Web. Os autores utilizam dois algoritmos no trabalho. O primeiro algoritmo, que verifica a semelhança entre os pares avaliados, variou sua acurácia entre 60%, para a desambiguação de verbos utilizando o melhor sentido; até 98% no caso de desambiguação de substantivos. No caso do segundo algoritmo, que utiliza o conceito de densidade semântica (distância semântica entre duas ou mais palavras), a acurácia obtida variou entre 67% e 97%. Nesse caso, a acurácia de 67% foi obtida na desambiguação de verbos e a acurácia de 97% para

a desambiguação de substantivos.

Banerjee and Pedersen (2002) apresentaram uma variação do algoritmo de Lesk (1986), substituindo os dicionários simples pelo *WordNet*. Nessa abordagem, a similaridade do seu conjunto de sentidos é computada para o seu conjunto de palavras similares, os *synsets*. O sentido com maior similaridade é adotado. Todo o demais funcionamento do algoritmo é baseado na proposta de Lesk (1986).

Li et al. (2012) propõe uma abordagem de não-supervisionada para desambiguação lexical. Esta é uma proposta que visa diminuir o custo computacional da desambiguação enquanto se obtém um bom valor para acurácia. Por utilizar modelos probabilísticos para a definição inicial dos parâmetros, a variação nos parâmetros conclui em resultados diferentes para a mesma entrada. Os testes dos autores variaram a acurácia entre 60% e 80%.

O trabalho de Nguyen and Ock (2013) utiliza algoritmos de otimização, adotando um grafo para realizar a tarefa de desambiguação de sentidos. Em sua proposta, os autores convertem o problema de desambiguação de sentidos em uma versão do Caixeiro Viajante. Na versão convertida, o grafo é composto pelos *synsets*, representados pelos vértices, e pela similaridade, representada pelas arestas ligando os *synsets*. O algoritmo insere o termo analisado no grafo, inicialmente desconectado dos demais vértices. Durante a execução, são adicionadas arestas ligando o termo analisado aos *synsets* presentes no grafo. O peso de cada aresta é calculado pela da métrica proposta por Lesk (1986). Após completar o grafo, é executado um algoritmo responsável por encontrar o menor caminho entre dois *synsets* e consequentemente o melhor sentido.

Wang et al. (2014) sugere uma abordagem supervisionada utilizando *semantic diffusion kernel*. Essa técnica cria um modelo semântico, por similaridade dos significados. O modelo é implementado em grafo definido pela co-ocorrência dos termos e significados, além de informações léxicas. Busca-se, assim, obter um modelo semelhante ao *bag of words*, em termo de representatividade de exemplos, contudo, com relações mais suaves. A proposta é superior a outras em diversos experimentos de desambiguação apresentados pelo *SensEval*.

Os trabalhos desta seção apresentam técnicas da literatura para desambiguação lexical de sentido. Algumas dessas técnicas são testadas por este trabalho, em busca do melhor algoritmo para ser utilizado como parte da metodologia de extração de conhecimento.

Capítulo 4

Manipulação das bases de dados para predizer o nível de conhecimento do usuário

Este capítulo apresenta a caracterização do usuário a qual ocorrerá a partir de seus interesses em conjunto com o seu nível de conhecimento. Para realizar essa caracterização, a primeira etapa consiste em gerar uma base de dados com diversas áreas de interesse e, em seguida, verificar o nível e dificuldade do texto, de acordo com a frequência dos termos utilizados. Neste capítulo, na Seção 4.1 é demonstrado como foi o processo de criação da base de dados. A Seção 4.2 elucida sobre o cálculo de dificuldade dos documentos coletados.

4.1 Bases de dados

A realização da etapa inicial necessitou da criação de duas bases de dados distintas. A primeira consiste em uma referência da frequência dos termos da língua inglesa. Essa base visa apresentar a regularidade de uso de cada um das palavras, em ambientes de escrita, seja jornais, revistas, legendas, músicas, entre outras fontes. A segunda armazena os dados referentes às possíveis áreas de interesse do indivíduo. A geração de cada uma é detalhada nas Subseções 4.1.1 e 4.1.2, respectivamente.

4.1.1 Base de Referência

Para a execução da metologia proposta, é necessário uma referência da frequência de uso das palavras de língua inglesa. De modo a obter esta informação, foi coletada uma base de referência, na qual é possível verificar a frequência de cada cada termo utilizado em ambientes impressos da língua inglesa. Os dados utilizados são disponibilizados pela *Gavagai*[1] em seu serviço *Lexicon*[2]. Essa é uma base consolidada, utilizada como referência para várias análises em que destaca-se Allan et al. (2012); Clements et al. (2008); Zuccon et al. (2014).

O *Lexicon* é responsável por ranquear os termos de língua inglesa dentro de uma coleção de 2,5 milhões documentos. Para cada termo é contabilizado o número de documentos em que ele está presente e o total de vezes de sua utilização na coleção. A partir disso, são calculados dois pesos para cada termo: absoluto e relativo. O peso absoluto expressa a representatividade do termo na base. A ordenação do vocabulário por esse peso gera *ranking* absoluto (*absolute ranking*) e explicita a frequência de um termo. Nela, o primeiro representa o mais frequente, seguido do segundo e assim sucessivamente.

O segundo peso é o relativo. A ordenação pelo peso relativo disponibiliza o *ranking* relativo (*relative ranking*) que atribui valores entre 0 (zero) e 1 (um) para cada termo da coleção. Nesse caso, valores mais próximos de zero indicam termos mais comuns da coleção e consequentemente mais fáceis. Em contrapartida, pesos que tendem a um indicam termos infrequentes e que apresentam maior dificuldade de uso e, possivelmente, de ser aprendido.

A partir desses fatos, para coletar a base de referência, foi implementado um serviço em linguagem *Java*[3] capaz de requerer as informações sobre cada termo e armazenar suas informações em banco de dados *PostgreSQL*[4].

As requisições ocorreram por demanda, ou seja, apenas os termos presentes na base que armazena os dados de interesse eram requisitados e armazenados para referência. Essa abordagem foi utilizada devido à necessidade de realização de uma requisição por termo, uma vez que esse é o único serviço disponível para coleta do *Lexicon*.

[1] https://gavagai.se/
[2] http://lexicon.gavagai.se/
[3] https://www.oracle.com/java/index.html
[4] http://www.postgreesql.org

4.1.2 Temas de Interesse

A base que armazena os temas de interesse foi gerada a partir de uma amostra dos vídeos disponibilizados no *YouTube*. Dentre os vídeos publicados nessa mídia digital, foram selecionados para compor a base aqueles que apresentavam as seguintes características: (1) o idioma de origem é o inglês; (2) contém legenda disponível para a língua inglesa. Vídeos com essas propriedades foram considerados aptos para comporem a base e entraram na fila para coleta. Dentre eles, foram identificados dois tipos de legendas: informadas pelo proprietário do vídeo e geradas automaticamente pelo sistema do *YouTube*. A qualidade dos subtítulos pode variar de acordo com o tipo do vídeo e o tipo da legenda disponível. Contudo, espera-se que aqueles informados pelo usuário tenham qualidade superior, transcrevendo exatamente o que é dito no vídeo.

Para compor a base, foram utilizadas legendas de vídeos do *YouTube*, independente do tipo associado. A criação utilizou uma biblioteca *Crawler4J*[5], implementada em linguagem *Java*. Esse coletor navega pelas páginas do *YouTube*, identifica aquelas que contenham vídeos que se adapta às características desejadas e armazena localmente as legendas encontradas. A escolha dessa ferramenta deve-se à familiaridade do autor com ela. Foram utilizadas sete categorias diferentes como sementes do coletor, a fim de obter conteúdo heterogêneo. A lista de categorias é apresentada na tabela 4.1

Tabela 4.1: Páginas utilizadas como semente.

Categoria
Página Inicial
Populares no YouTube
Música
Esportes
Jogos
Filmes
Notícias

Cada vídeo identificado de acordo com as características pré-determinadas foi coletado. Para isso, seu formato original utilizado pelo *YouTube*, *eXtensible Markup Lan-*

[5]https://github.com/yasserg/crawler4j

guage (XML[6]), foi convertido para *SubRip*[7] (extensão .srt). Além das legendas, outras informações também foram identificadas e armazenadas no processo de coleta. Categoria, duração do vídeo, data de publicação, nome do canal ao qual o vídeo está integrado e identificador do vídeo também foram coletados.

A fim de se obter os dados estatísticos dos vídeos, foi realizada a implementação de um serviço que realiza requisições *JavaScript Object Notation* (JSON[8]) para obter esses dados. A partir desse serviço, obteve-se: o número de *likes*, *dislikes* e visualizações dos vídeos. Esses dados foram armazenados em um banco de dados *PostgreSQL* de modo a facilitar a sua posterior recuperação. O tempo de coleta foi de aproximadamente trinta segundos por legenda. Isso ocorreu devido à necessidade de realizar três requisições aos servidores do *YouTube* para cada vídeo a ser coletado. A primeira para verificar se o vídeo cumpre as exigências necessárias. Caso positivo, eram acessados dois serviços distintos: um responsável por armazenar a legenda e outro que contém os demais dados do vídeo.

Após a geração da base para análise, tornou-se necessário desenvolver um método para facilitar a identificação dos dados a serem analisados. Para isso, foi realizado um processamento dividido em duas fases. A primeira fase converte as legendas para o formato *Comma Separated Values* (CSV[9]). Nesse novo formato, cada linha da legenda representa unicamente uma frase, diferentemente do formato *SubRip* no qual são necessárias pelo menos quatro linhas. No CSV, cada linha contempla as três informações características do *SubRip*: o identificador da linha, os tempos de início e fim da frase e o texto. Essa conversão gera uma base simplificada para a segunda fase do processamento.

A segunda fase consiste na indexação das legendas. Foi utilizado o *framework Lucene*[10] nessa etapa. A indexação visa simplificar a recuperação dos termos presente na base. Por se tratar de um conjunto de dados não estruturado, essa técnica de RI apresenta maior eficiência para obter as informações armazenadas. O processo de indexação armazenou separadamente a linha referente à frase, o tempo de início, o tempo de final, a duração da exibição da frase, o número de palavras contidas na frase e o número de palavras por segundo ditas na frase. Além dessas informações, é possível recuperar, também, as frequências do termo no documento (TF), na coleção (DF) e a frequência

[6] http://www.w3.org/XML
[7] http://sourceforge.net/projects/subrip/
[8] http://www.json.org
[9] http://creativyst.com/Doc/Articles/CSV/CSV01.htm
[10] http://lucene.apache.org/

invertida na coleção (IDF).

4.2 Caracterização do Nível de Dificuldade dos Documentos

A primeira amostragem que foi analisada contou com 35.000 legendas de vídeos do *YouTube*. Destas, foram extraídos um total de 90.384 termos, sendo 293 *stopwords*. Uma vez coletados os dados, iniciou-se o processo de dar peso às legendas de acordo com a sua dificuldade. Para isso, as seguintes premissas foram consideradas:

- Quanto menor a frequência do termo na base de referência, maior a dificuldade da frase na qual o termo está inserido;
- O número de termos varia de acordo com o tamanho da legenda e esse fato não pode interferir no resultado.

Os pesos são calculados segundo a Equação 4.1, em que: P corresponde ao resultado da equação e representa o peso da legenda, N é o número *tokens*, i representa o i-ésimo termo da legenda, v é o número de termos da legenda; rr é o *ranking* relativo (*relative rank*) do termo, obtido por meio do *Lexicon* e representa o peso do termo dentro da língua; tf_i, consiste na frequência do i-ésimo termo da legenda.

$$P = \frac{1}{N} * \sum_{i=1}^{v} rr * tf_i \tag{4.1}$$

Outras fórmulas foram testadas, mas a citada demonstrou ser mais efetiva em relação às demais. Como resultado dessa equação, obtém-se o peso da legenda ponderando pelo número de *tokens*. A Tabela **??** apresenta os 10 vídeos considerados mais difíceis, de acordo com a métrica proposta na Equação 4.1. Observe que há vídeos com apenas treze *tokens* e vídeos com milhares de *tokens*. Destaca-se, ainda, que, na coluna "Sem *stopwords*", é possível observar a posição em que os vídeos destacados são alocados ao desconsiderar as *stopwords* presentes neles. A partir desses dados optou-se por trabalhar sem a remoção das *stopwords*. Isso ocorreu porque, durante os estudos, todos os termos podem ser de interesse do aluno.

,,

,,,,

Capítulo 5

Metodologia de Extração de Conhecimento para Geração de Exercícios

Este capítulo visa apresentar uma metodologia para possibilitar o aprendizado de vocabulário de língua inglesa a partir de temas de interesse do usuário. Essa metodologia visa ponderar a relação entre dois termos que são sinônimos e, assim, gerar uma base de conhecimento. A partir da base de conhecimento, juntamente com documentos informados pelo usuário, gerar exercícios automaticamente.

A Figura 5.1 apresenta o fluxograma da metodologia proposta, para possibilitar o aprendizado de vocabulário da língua inglesa a partir de temas de interesse do usuário. Verifica-se a presença de dois componentes. O primeiro representa o Extrator de Conhecimento, componente responsável por processar os dados contidos em um dicionário, de modo a ponderar as relações de sinonímia presentes e armazenar em uma Base de Conhecimento. O processo de extração de conhecimento utiliza duas bases de dados: a *WordNet*, que representa o dicionário no qual são extraídos os dados para se identificar o peso da sinonímia entre dois termos; e a Base de Conhecimento, que é a base de dados na qual as relações e pesos identificados durante o processo são armazenados.

O segundo componente, também apresentado na Figura 5.1, representa o Gerador de Exercícios. Ele é responsável por gerar exercícios de vocabulário a partir de um documento, de entrada, informado pelo usuário. A Base de Conhecimento, gerada no Extrator de Conhecimento, representa a outra entrada para esse componente. Os exercícios

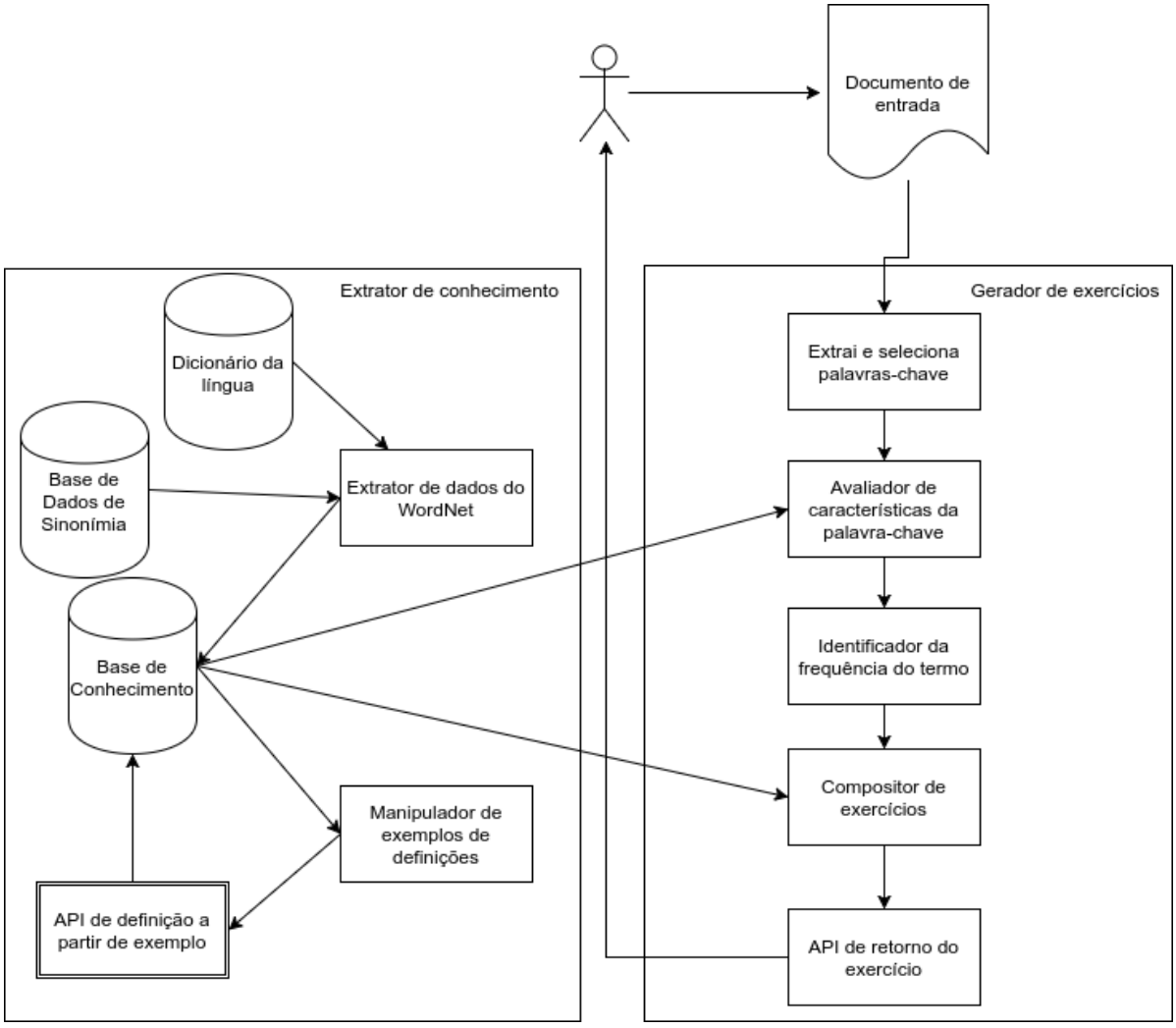

Figura 5.1: Fluxograma da Metodologia de Extração de Conhecimento para Geração de Exercícios.

gerados são enviados ao usuário para que possam ser respondidos.

Conforme supracitado, a metodologia é dividida em dois componentes e cada um deles é detalhado no decorrer deste capítulo. Inicialmente, na Seção 5.1, é apresentado o componente para Extração de Conhecimento. Em seguida, a Seção 5.2 detalha o componente Gerador de Exercícios.

5.1 Extrator de Conhecimento

O Extrator de Conhecimento é o componente responsável por ponderar a relação de sinonímia presente entre pares de termos. Partindo do conceito apresentado, é possível inferir que duas palavras que são sinônimas apresentam um grau de similaridade, a qual pode gerar sinônimos perfeitos ou imperfeitos (Houaiss; 2003). É dito sinônimo perfeito quando dois termos distintos têm significados idênticos. Sinônimos imperfeitos são aqueles termos que têm significados próximos, mas não idênticos.

O Extrator de Conhecimento parte da premissa de que dois termos listados como sinônimos são imperfeitos e que é possível estimar o grau de sinonímia entre eles. O grau de sinonímia é calculado a partir da similaridade das definições utilizadas, de ambos os termos, em determinada frase. Logo, a similaridade identificada entre o par de termos será definida pela relação de uma definição de cada um dos termos. Por exemplo, a partir dos termos '*sense*' e '*feel*', que são sinônimos, baseando-se na *WordNet*, será possível definir que a definição de '*sense*': *a faculty by which the body perceives an external stimulus* tem uma maior similaridade com o termo '*feel*' para o significado 1 (*an act of touching something to examine it*), do que para o significado 2 (*experience (an emotion or sensation)*).

Os componentes que formam o Extrator de Conhecimento são apresentados a seguir. Inicialmente, a Seção 5.1.1 apresenta a Base de dados de Sinonímia utilizada como entrada para o Extrator de Conhecimento. Em seguida, a Seção 5.1.2 apresenta o Extrator de Dados da *WordNet*, que populará a Base de Conhecimento. A Seção 5.1.3 especifica o Manipulador de Exemplos de Definições. Em seguida, a Seção 5.1.4 apresenta a API de Definição a partir de exemplo, que irá complementar a Base de Conhecimento.

5.1.1 Base de Dados de Sinonímia

Iniciar o Extrator de Conhecimento necessita de uma base de dados de sinonímia e de um dicionário da língua inglesa. A base de dados de sinonímia é utilizada para identificar os sinônimos da palavra processada. O dicionário da língua, conforme apresentado na Seção 2.2.2, contém as definições das palavras, uma das informações utilizadas nessa metodologia.

A base de dados de sinonímia utilizada é a *WordNet*, por ser uma base de dados léxica amplamente utilizada em trabalhos relacionados ao PLN, conforme apresentado

na Seção 2.2.4. Outro fato que viabiliza o uso da *WordNet* é a presença de um dicionário em sua base de dados. Assim, os dados utilizados pelo Extrator de Conhecimento são obtidos a partir de uma única fonte de dados.

5.1.2 Extrator de Dados da WordNet

De modo a iniciar o Extrator de Conhecimento, é necessário extrair os dados necessários da base de dados de sinonímia. O conjunto de dados necessários é composto pela lista de termos da língua inglesa e sua coleção de definições, juntamente com a relação de sinonímia entre os termos. Conforme supracitado, esses dados estão disponíveis na *WordNet* e ela é a base de dados utilizada como entrada nesse processo. A partir dela, é necessário que sejam extraídos os dados relevantes para a metodologia. Logo, o Extrator de Dados da *WordNet* é responsável por extrair essas informações e gerar a Base de Conhecimento.

A extração é realizada em duas etapas distintas que se complementam. A primeira etapa é responsável por coletar cada termo, suas definições e exemplos de uso de cada definição. A execução necessita da realização de uma consulta sequencial na base. Nessa consulta, para todo termo presente na *WordNet* é extraído o seu conjunto de definições e para cada definição é extraído o conjunto de exemplos de uso. Todos esses dados são armazenados na Base de Conhecimento.

A Figura 5.2 apresenta a estrutura dos dados extraídos. Nela, é possível visualizar que um termo presente na *WordNet* contém uma estrutura de definições associadas a ele. Em seguida, existe uma estrutura composta por frases, que são exemplso de uso do termo, para aquela definição. Essa cadeia de dados é extraída e armazenada na Base de Conhecimento.

A segunda etapa realizada pelo Extrator de Dados da *WordNet* condiz com a identificação de sinonímia dos termos. Novamente, a sua execução necessita de uma consulta sequencial na base de dados da *WordNet*. Entretanto, nesse ponto, a consulta é realizada pelos *synsets*, ou agrupamentos de termos sinônimos. Logo, ao buscar um termo, são identificados aqueles termos que estão presentes no mesmo *synset*. Todos os termos encontrados são coletados e a relação do termo com seus sinônimos é armazenada na Base de Conhecimento. Por exemplo, ao buscar os sinônimos do termo '*feel*', é obtida a lista: *experience*, *find*, *sense*.

Ao finalizar esse processamento das duas etapas do Extrator de Conhecimento, está

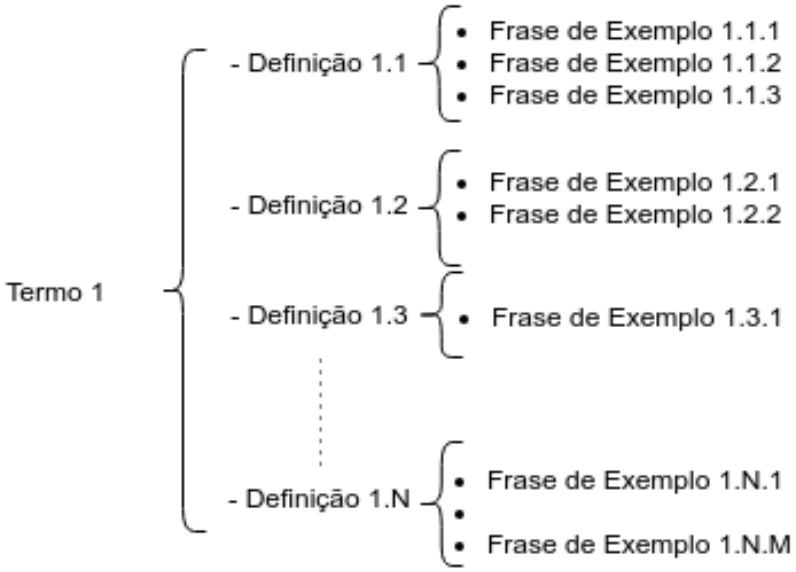

Figura 5.2: Exemplo de estrutura dos dados extraídos.

formada uma relação de sinônimos diferente daquela presente na *WordNet*. Isso ocorre porque, na *WordNet*, os sinônimos são armazenado como pertencentes ao termo, ou seja, não há diferenciação entre os sinônimos. Na base de conhecimento, as informações são armazenadas em função das definições, por exemplo, as definições de '*feel*' são sinônimas das definições de '*sense*'. Esta modelagem complementa a estrutura atual da *WordNet*. Com isso, a base de termos possibilita um número maior de comparações entre dois termos, abrangendo a quantidade de dados e, consequentemente, a variação nos resultados a serem obtidos.

Nesse momento, a Base de Conhecimento é uma base de dados composta por termos e suas definições, juntamente com a relação de sinonímia dos termos, baseando-se nas definições. Outros dados disponíveis na *WordNet* poderiam ser extraídos nessa fase, entretanto, a sua disposição original possibilita a continuidade do processo e, logo, de adicionar mais dados na Base de Conhecimento. Concluída a execução do Extrator de Conhecimento da *WordNet*, está formada a versão básica da Base de Conhecimento. Entretanto, mesmo existindo a relação de sinonímia entre os termos, ainda não há ponderação da sinonímia para as definições. De modo a iniciar esse cálculo, é necessário executar o Manipulador de Exemplos de Definições.

5.1.3 Manipulador de Exemplos de Definições

Conforme supracitado, o cálculo da similaridade entre dois termos que são sinônimos é realizado a partir das definições dos termos. A associação direta entre definições não é uma tarefa simples, pois devem ser consideradas muitas informações de difícil computação, como: o contexto de uso do termo, palavras próximas ao termo analisado, o tempo verbal da frase, entre outras informações.

De modo a diminuir a computação para identificar o contexto de ambos os termos, optou-se por utilizar exemplos de uso das definições de dicionário. Esses exemplos de uso são frases nas quais o termo é utilizado. O ponto importante é que o exemplo de uso está associado a uma definição, ou seja, o termo naquela frase tem uma definição específica associada ao termo e esta definição já é previamente conhecida. Essa estrutura, apresentada na Figura 5.2, é a estrutura na qual a Base de Conhecimento é construída. A partir dela, o Manipulador de Exemplos de Definições pode ser executado.

O Manipulador de Exemplos de Definições é o componente responsável por iniciar o processo de calcular a similaridade entre pares de termos sinônimos. Sua execução consiste na formação da frase na qual será realizada a análise da similaridade entre cada par de termos. Para isso, a partir da Base de Conhecimento, são selecionados os seguintes dados: o termo analisado, uma definição do termo, o sinônimo ao qual se deseja calcular o grau de similaridade com o termo analisado, uma das definições do sinônimo e a lista de exemplos de uso da definição do sinônimo.

Em seguida, o sinônimo, presente na frase de exemplo, é substituído pelo termo sob análise. Assim é formada uma nova frase, semelhante à primeira, mas com a alteração de um dos *tokens*. Para exemplificar, consideremos que será analisada a similaridade do termo ***sense*** (sensação) com seu sinônimo ***feel*** (sentir), quando o seu sentido é: *an intuitive awareness* (uma consciência intuitiva). A Figura 5.3 apresenta como a operação é realizada. Inicialmente, a definição *an intuitive awareness* contém duas frases de exemplo, com o termo *fell*: *he has a feel for animals* e *it's easy when you get the feel of it*.

De modo a ser processado pela API de definição a partir de exemplo, o termo *feel* é substituído por *sense* em ambas as frases. Assim, as novas frases, conforme observado na Figura 5.3, foram: *he has a sense for animals* e *it's easy when you get the sense of it*. Ao realizar a substituição, de modo a manter a coerência da frase, são trabalhados sempre termos e definições com as mesmas funções sintáticas. Ou seja, é verificada a

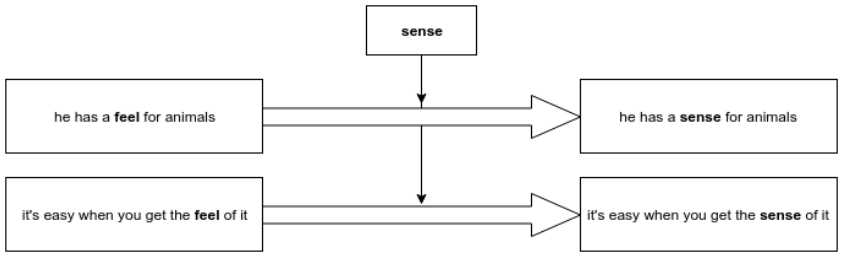

Figura 5.3: Exemplo da substituição do termo pelo seu sinônimo na frase de exemplo.

relação entre a definição de dois verbos ou dois substantivos, mas nunca entre um verbo e um adjetivo. A substituição do sinônimo pelo termo sob análise busca identificar quanto cada definição do termo se assemelha à definição analisada do sinônimo. Essa comparação é possível devido a dois fatores: (1) o exemplo de uso está diretamente relacionada a uma definição do sinônimo; (2) é possível calcular qual definição, do termo sob análise, se aplica à frase alterada. Essa análise é realizada pela API de Definição a partir de Exemplo.

5.1.4 API de Definição a Partir de Exemplo

A API de Definição a partir de Exemplo trata-se do componente final da metodologia do Extrator de Conhecimento. A sua execução recebe como entrada os dados de saída do Manipulador de Exemplos de Definição. Sendo eles: o termo a ser analisado, o sinônimo, a definição do sinônimo e as frases de exemplo, do sinônimo, modificadas. Por exemplo, ao analisar o termo '*feel*' e seu sinônimo '*sense*', a API de Definição a partir de Exemplo recebe por parâmetro: o termo '*feel*', o sinônimo sob análise '*sense*', a definição de '*feel*' (*experience (an emotion or sensation)*) e, por fim, as frase que exemplificam esta definição, já modificadas pelo Manipulador de Exemplos de Definições: *he has a* ***sense*** *for animals* e *it is easy when you get the* ***sense*** *of it*.

Como nesse momento as frases a serem analisadas já foram manipuladas, torna-se necessário identificar o peso que cada definição do novo termo, presente nas frases de exemplo, tem para o contexto em que está sendo utilizado. A identificação dos pesos associados aos sentidos do termo utilizado dentro de determinado contexto é parte da pesquisa de Desambiguação Lexical de Sentido (DLS). O PyWSD, *Python implementation for Word Sense Disambiguation*, desenvolvida por Tan (2014), é a implementação na qual foi baseado esse serviço. A escolha ocorreu devido ao número considerável de

propostas que estão implementadas na ferramenta, dentre elas Lesk (1986), Leacock and Chodorow (1998), Banerjee and Pedersen (2002), Banerjee and Pedersen (2003) e Lee et al. (2004).

Todas as implementações supracitadas utilizam como entrada: uma frase e o termo da frase a ser analisado. A execução de um sistema DLS busca identificar, para o dado termo, qual de suas definições é a mais provável de ser corretamente associada àquele contexto. A melhor definição encontrada é a saída do método.

Dentre as diversas implementações presentes na biblioteca, a proposta por Banerjee and Pedersen (2003) adaptada para calcular similaridades por cosseno (*cosine lesk*) foi selecionada para ser utilizada nesse serviço. Como a implementação busca identificar apenas qual das definições de um termo é utilizada no contexto, tornou-se necessária uma adaptação para o caso de uso em que ela será utilizada.

Na versão adaptada, uma vez calculados os pesos de todas as definições para o termo processado, é formado um conjunto de saída que contenha a definição e o peso que se refere ao quanto aquela definição se adapta àquele contexto. Esse conjunto é retornado em ordem decrescente, de modo que o primeiro par definição/peso tenha maior peso que o segundo, e o segundo seja maior que o terceiro e assim sucessivamente.

Com isso, é possível identificar o sentido de um termo dentro de dado contexto, logo, é iniciada a API de definição a partir de Exemplos. A execução condiz com a execução do PyWSD para identificar o sentido do termo utilizado no contexto.

Conforme citado anteriormente, é utilizada a implementação de Banerjee and Pedersen (2003) para calcular a probabilidade de cada sentido. O método *cosine lesk* adaptado, presente no PyWSD, é invocado, utilizado como parâmetro a frase, cujo sinônimo foi substituído pelo termo, e o termo a ser processado. Finalizada a execução, é gerada uma lista com as definições do termo original, ordenadas pelo peso calculado.

A lista resultante representa o peso associado a cada definição do termo na frase. Como a frase na qual o termo está presente é um exemplo de uso do sinônimo, obtém-se, assim, o peso que associa as definições de dois termos distintos. Como o número de frases de exemplo varia entre as definição, o peso final da relação é obtido por meio da média dos pesos da definição para todas as frases de exemplo.

A Figura 5.4 demonstra o resultado parcial da execução da API de Definição a partir de Exemplo. Nela, é possível visualizar que para cada frase existe uma lista de definições do termo *feel*. Entretanto, em cada uma das frases o peso associado é diferente. Os pesos

Sentido	Peso
feel.n.1	0,02380
feel.n.2	0,01306
feel.n.5	0,001055

Tabela 5.1: Pesos das definições de *feel* para uma definição de *sense*.

definidos às definições variam entre zero e um, onde zero indica que não há qualquer similaridade entre os termos analisados, enquanto um sugere uma grande similaridade. Baseando-se unicamente no exemplo presente na Figura 5.4, a relação entre o sentido *an intuitive awareness* do termo *sense* e os diversos sentidos do termo *feel* seria representada pela Tabela 5.1. Essa tabela é composta por duas colunas, nas quais: a primeira, denominada Sentido, apresenta um código, utilizado pela *WordNet*, para representar as definições dos termos; a segunda, denominada Peso, apresenta o peso gerado pelo algoritmo para cada definição. Ainda nessa tabela, é possível visualizar que, na média, o sentido *feel.n.1* é o que mais se aproxima do sentido analisado de *sense*, fato demostrado pelo seu peso ser superior aos demais sentidos.

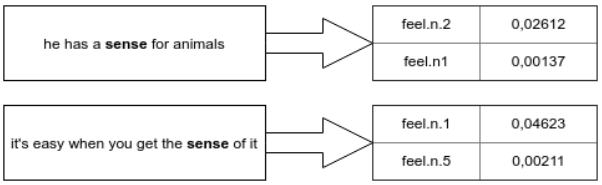

Figura 5.4: Exemplo de pesos associados a sentidos em contextos diferentes.

Finalizado o processo, os dados obtidos são incluídos na Base de Conhecimento, atualizando, assim, a relação de sinonímia entre termos anteriormente presentes nessa base. Neste momento, além da associação entre as definições, está presente também o peso de cada sinônimo. O componente realiza a sua execução sempre que é acionado pelo Manipulador de Exemplos de Definições, logo, a execução de ambos continua, em laço, até que não existam mais relações de sinonímia sem pesos associados. A Tabela 5.2 exemplifica a base de conhecimento neste momento. Nela é possível observar que duas definições do termo *sense* foram ponderadas, como sinônimos, em relação ao termo *feel*. Ainda na Tabela 5.2 é possível observar que a definição *sense.n.1* tem maior similaridade com *feel.n.1*, enquanto *sense.n.2* se assemelha mais com *feel.n.5*. Nesse ponto é possível

Termo	Definição	Sinônimo	Definição	Peso
sense	sense.n.1	feel	feel.n.1	0,02380
sense	sense.n.1	feel	feel.n.5	0,00105
sense	sense.n.2	feel	feel.n.1	0,01406
sense	sense.n.2	feel	feel.n.5	0,02212

Tabela 5.2: Visão da base de conhecimento após a ponderação a partir de exemplo

observar que dois termos, definidos como sinônimos na *WordNet*, tem similaridades diferentes de acordo com o contexto que estão sendo utilizados.

Com o finalizar do processamento, temos a relação entre cada termo com todos os seus sinônimos para cada definição do termo original. Entretanto, o peso associado pode variar, demostrando uma maior ou menor relação entre os termos. Fato também apresentado na Tabela 5.2. Assim, é possível definir a relação entre dois termos a partir do peso das relações entre as suas definições. Logo, quanto maior o peso das relações entre as definições, maior a relação entre os termos aos quais pertencem as definições.

Conforme citado anteriormente, os pesos para as associações são variáveis. O número de definições dos termos também é inconstante. Logo, para utilizar da relação entre os termos é necessário realizar a normalização entre os dados disponíveis. Para isso, foram realizados estudos de algumas propostas. A primeira proposta é representada pela equação $p = (\sum_{i=0}^{n} p_i)/n$, na qual i indica o i-ésimo termo sob análise, n é o número de definições do termo e p_i é o peso identificado para a definição i. Essa equação representa a ponderação do termo de acordo com a média dos pesos das definições. Com isso, é possível definir qual a relação média entre dois termos.

Uma segunda proposta condiz com a identificação da facilidade de relacionar uma definição a determinado contexto. Com essa informação é possível, por exemplo, identificar a dificuldade de um exercício. Outro fato é a possibilidade de determinar o quão diferentes duas definições são. Para isso, é calculada a diferença entre o peso da melhor definição e o peso de todas as demais definições. Para essa proposta, foi necessário determinar dois limiares: (1) o limiar de similaridade, que determina que duas definições são matematicamente iguais por a diferença entre elas ser menor que esse fator; (2) o limiar de dificuldade, que identifica a dificuldade de gerar exercícios confiáveis com essa

relação.

O limiar de similaridade deve ser, a princípio, um valor baixo. Isso ocorre porque a relação entre os pesos e a similaridade é diretamente proporcional. Em contrapartida, o limiar de dificuldade deve ser um número mais alto. Isso ocorre devido ao mesmo princípio de proporcionalidade. A partir da definição desses fatores, o conjunto resultante é dividido em três grupos: (1) o grupo de maior similaridade, em que todas as definições têm alta relação com a definição trabalhada; (2) o grupo intermediário, no qual podem haver definições do sinônimo que se relacionam ou não com a definição a ser trabalhada, de todo modo, essa separação não é clara e (3) o grupo de menor similaridade, em que a similaridade tende a zero e consequentemente há menores relações com com a definição trabalhada.

A definição dos limiares é uma tarefa complexa. Para este trabalho foram testados vários valores a fim de encontrar a melhor relação. Os testes realizados são apresentados no Capítulo 6.

5.2 Gerador de Exercícios

A fim de utilizar o conhecimento extraído da *WordNet* por meio do Extrator de Conhecimento e armazenado na Base de Conhecimento, foi proposto um segundo componente pertencente à Metodologia de Extração de Conhecimento e Geração de Exercícios. Este componente é o Gerador de Exercícios.

Tal componente é responsável por gerar exercícios a partir de documentos de entrada informados pelo usuário. Juntamente com esse documento, é utilizado o conhecimento previamente obtido, a partir do Extrator de Conhecimento.

O componente de Geração de Exercícios é composto por diversos subcomponentes. Inicialmente, a Seção 5.2.1 apresenta o Extrator de Palavras-Chave, responsável por selecionar as palavras utilizadas para gerar os exercícios. Em seguida, a Seção 5.2.2 demonstra o funcionamento do Avaliador de características da palavra-chave, componente no qual é avaliada a possibilidade de geração de exercícios a partir da palavra-chave. Complementarmente, a Seção 5.2.3 identifica a frequência de uso do termo na língua inglesa. A Seção 5.2.4 apresenta o Gerador de Exercícios, responsável por gerar os exercícios que utilizam a palavra-chave. Por fim, a Seção 5.2.5 apresenta a API de Retorno do Exercício, componente responsável por retornar os exercícios gerados para o

usuário solicitante.

5.2.1 Extrator de Palavras-Chave

De modo a iniciar o Gerador de Exercícios, inicialmente é necessário selecionar e extrair os termos que serão utilizados para gerar os exercícios. Esses termos são extraídos a partir do documento de entrada informado pelo usuário. Os termos utilizados na geração de exercícios são chamados de palavras-chave, por serem os itens centrais no processo de geração de exercícios.

O Extrator de palavras-chave é o componente responsável por identificar e extrair os termos utilizados para a geração dos exercícios. De modo a iniciar a sua execução, primeiramente, é realizado um pré-processamento no documento de entrada. Nesse pré-processamento, o documento foi normalizado de modo a possibilitar uma extração uniforme das palavras-chave. A normalização consiste em realizar o *stemming* nas palavras, de forma que seja possível identificar qual o radical formador das palavras do documento.

O pré-processamento consistiu na conversão do documento de entrada em um conjunto de frases. A identificação das frases foi realizada a partir da pontuação utilizada na escrita das mesmas. Logo, para o documento de entrada, sempre que é identificado um sinal de pontuação, seja ponto final, exclamação ou interrogação, uma nova frase é iniciada a partir da próxima palavra. A Figura 5.5 ilustra este processo. Nela é possível visualizar que um documento foi repartido em três frases

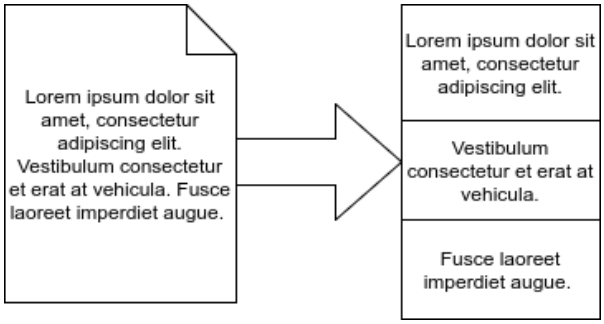

Figura 5.5: Exemplo de pesos associados a sentidos em contextos diferentes.

Finalizada a divisão do documento em frases, é iniciada a fase de seleção das palavras-chave. Nesse seleção foram consideradas palavras-chave apenas verbos, substantivos e adjetivos. Esses tipos de termos foram selecionados devido, principalmente, à maior

representatividade de significado desses termos nas frases. Definidos os termos a serem extraídos, tornou-se necessária a identificação desses termos nas frases. Para realizar a identificação do tipo de cada palavra, é realizado pelo *POS Tagger* disponibilizado pela *Stanford*[1].

A execução do *POS Tagger* identifica os termos que são candidatos a serem palavras-chave. A indicação ocorre através da rotulagem de todos os *tokens* presentes na frase. O rótulo indica qual a função de cada *token* na frase. Os pré-selecionados como palavra-chave são aqueles rotulados como verbos, substantivos e adjetivos. Para determinar se ele será considerado uma palavra-chave e, consequentemente, se serão gerados exercícios utilizando-o, é necessário avaliar se o termo tem algumas características importantes. Para isso, foi utilizado o componente Avaliador de características da palavra-chave.

5.2.2 Avaliador de Características da Palavra-Chave

O Avaliador de Características da Palavra-Chave é o componente responsável por validar se é possível a geração de exercícios a partir da frase. Este componente recebe como entrada a frase rotulada do documento de entrada. Entretanto a avaliação ocorre somente para os termos pré-selecionados como palavras-chave. A validação ocorre a partir de características que são observadas nas palavras-chaves.

Inicialmente, busca-se identificar se as definições do termo são significativamente diferentes para serem consideradas como opções de um exercício. Este dado é observado ao se obter os pesos definições das palavras-chave dentro do contexto. Para essa tarefa, novamente é utilizado o desambiguador *PyWSD*. Neste processo, busca-se retirar do processo de geração de exercícios, os termos que o desambiguador ponderar todas as definições com valores inferiores a determinado limiar. Esse limiar é responsável por identificar as melhores definições daquelas que não são satisfatórias, em determinado contexto de uso do termo. A definição do limiar utilizado para a geração de exercícios foi realizada a partir de análises de diversos valores.

Com o limiar definido, o número de definições utilizadas para a geração de exercícios é limitado, sendo consideradas apenas aquelas cujo peso respeita o limiar definido. A melhor definição, de acordo com o peso, é selecionada. As definições cujo peso seja menor do que a diferença entre o peso da melhor definição e o limiar, segundo a fórmula $p_i <= p_1 - lim$, também são selecionadas. Outras características são utilizadas pelo

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

Avaliador de Características da Palavra-Chave, sendo elas: o número de definições do sentido indicado como correto, o total de frases de exemplo da palavra, o número de exemplos do sentido indicado como correto e a soma de todos os pesos associados ao termo. O modo de uso dessas características é descrito na Seção 6.4.

Se houver menos de quatro definições após a avaliação do limiar, a frase é descartada e não são gerados exercícios com ela. Caso haja exatamente quatro definições, uma nova definição é buscada para complementar a lista de opções do exercício. Essa nova definição é o melhor sinônimo da pior definição da palavra-chave avaliada. A informação é obtida a partir de uma consulta na Base de Conhecimento.

Caso existam cinco definições, essas serão utilizadas para a geração do exercício, não sendo necessária a execução de nenhuma tarefa adicional de seleção. Por fim, se houver mais de cinco definições no subconjunto resultante, a melhor definição é selecionada e a composição das outras quatro será realizada aleatoriamente. A aleatoriedade é utilizada uma vez que, a partir do conceito da utilização do limiar, todas as definições obtidas para determinado limiar previamente definido podem ser consideradas equivalentes.

A última característica utilizada na geração dos exercícios consiste na identificação da frequência de uso do termo dentro da língua inglesa.

5.2.3 Identificador da Frequência do Termo

A identificação da dificuldade de um exercício é realizada pela popularidade do termo utilizado como base para a geração do exercício. De modo a utilizar esses dados, foi necessária a geração de uma base de termos com as suas respectivas frequências de uso. A obtenção desses dados demanda de uma coleção grande de dados, com diversos textos coletados de fontes distintas, assuntos diversos, de modo a obter dados que representem a realidade do idioma.

A geração de uma base como essa demanda tempo e recursos que poderiam diminuir o foco em outras partes igualmente relevantes para a realização deste trabalho. Por isso, buscou-se identificar uma base de dados que suprisse essa necessidade. A base de dados selecionada e utilizada neste trabalho é o *Gavagai*, já detalhada na Seção 4.1.1.

Conforme citado no Capítulo 2, Breland (1996) relaciona a frequência de uso de um termo em texto do idioma com a dificuldade desse termo. Seu trabalho indica que termos frequentes são mais fáceis do que termos infrequentes. Logo a aplicação desses termos

alteram a complexidade de uma frase. Com isso, é possível variar a dificuldade dos exercícios de uma mesma frase selecionando termos diferentes para basear o exercício. De modo a possibilitar essa dificuldade variável, é utilizado o *Lexicon* como serviço que identifica a frequência de uso dos termos da língua inglesa.

Para cada termo selecionado no avaliador de característica, é atribuído a ele um peso relativo, correspondente ao *relative ranking* disponibilizado pelo *Lexicon*. O peso é obtido através da equação $(\sum_{i=0}^{n} r_i)/n$, na qual i representa o i-ésimo *token* da frase, n é o total de *tokens* da frase e, r_i é a frequência relativa do i-ésimo *token*. A partir desse peso, após a geração dos exercícios, torna-se possível ordenar os exercícios gerados de acordo com a dificuldade. Pesos menores indicam frases mais simples em contrapartida a pesos maiores que condizem com uma maior complexidade da frase.

Finalizadas as avaliações de todas as características relevantes do termo e a associação do peso aos termos, é iniciada a geração dos exercícios.

5.2.4 Compositor de Exercícios

O gerador de exercícios é o componente responsável por unir os dados de entrada informados pelo usuário e o conjunto de informações previamente processadas no extrator de conhecimento e armazenadas na base de conhecimento. Os exercícios gerados por esse componente são totalmente baseados no documento de entrada informado pelo usuário e são totalmente automatizados, ou seja, são gerados e corrigidos automaticamente.

Os três exercícios gerados são: o significado de um termo dentro de determinado contexto; os significados de um termo e qual sinônimo pode substituir um termo em determinado contexto. Todos os exercícios gerados são de múltipla escolha e utilizam conhecimentos extraídos a partir do extrator de conhecimentos.

A cada exercício gerado é atribuído um peso, que é representado pela média dos pesos das opções selecionadas para a geração desse exercício. Esse peso é utilizado como um guia para a dificuldade do exercício, pois também se utiliza do limiar para ser calculado. Logo, pesos maiores representam exercícios mais fáceis, pois o conjunto de opções é composto por itens mais distantes do limiar e, consequentemente, da opção correta.

Todos os exercícios seguem a mesma estrutura: são questões de múltipla escolha, compostas por uma opção correta e quatro opções incorretas. O gerador utiliza as informações previamente obtidas e armazenadas na base de termos e disponibiliza um

exercício de cada tipo.

Significado de um Termo em Contexto

O primeiro tipo de exercício gerado foi a identificação do significado de determinado termo para um contexto. A pergunta, referente a esse exercício é: "Qual dos significados listados condiz com o uso da palavra *termo* na frase *contexto*.", onde *termo* representa a palavra-chave utilizada na geração do exercício e *contexto* é a frase do documento de entrada informado pelo usuário.

A lista de opções é composta por um conjunto, pré-selecionado, de definições do termo, conforme supracitado. Não são selecionadas novas opções para compor a lista de exercícios e, consequentemente, as opções provenientes do componente anterior são utilizadas como opções do exercício.

Definições de um Termo

O exercício de identificação do significado de um termo, busca que o usuário responda a pergunta: "Qual dos significados abaixo listados é referente ao termo *termo*." A lista de opções é composta por um conjunto de definições.

Diferentemente do exercício anterior, a lista de opções utilizada nessa geração não e exclusivamente proveniente do componente anterior. A opção correta é a mesma anteriormente avaliada, contendo, assim, a definição do termo utilizado no exercício. Contudo, as opções incorretas são buscadas a partir da lista de sinônimos armazenada na base de termos.

A seleção das definições a serem utilizadas inicia-se com a identificação dos sinônimos que serão utilizados na geração do exercício. São selecionados aqueles sinônimos que, de acordo com as relações identificadas, não são sinônimos próximos, na média, do termo original. São selecionados os dois sinônimos que, na média, têm menor semelhança com o termo original. São selecionados sinônimos distantes, na média, pois o exercício não trabalha com o termo em um contexto específico e sim com uma definição qualquer do termo.

Uma vez definidos os termos, são selecionadas as definições daqueles termos que serão utilizados como opções. Como, inicialmente, a relação de sinonímia dos termos é genérica, são selecionadas duas definições quaisquer de cada um dos sinônimos para

compor o conjunto de opções.

Substituição de Termo por Sinônimo

A substituição de termo por sinônimo é o último tipo de exercício gerado. Ele busca identificar qual o sinônimo pode substituir um termo em determinada frase sem alterar o sentido original da frase. A pergunta é elaborada da seguinte forma: "Qual dos termos pode substituir a palavra *termo* na frase *contexto*, sem alterar o sentido da frase?"

Nesse caso, o *termo* representa a palavra-chave que será substituída por uma das opções listadas. Já o *contexto* representa o exemplo de uso da palavra-chave dentro do documento de entrada informado pelo usuário. Para selecionar as opções, é utilizado o conhecimento anteriormente adquirido de relacionamento entre sinônimos.

Para a opção correta, é identificado o melhor sinônimo do termo. No caso de opções incorretas, são buscados sinônimos do termo que não tenham uma relação forte para o contexto específico. O cálculo de distância baseia-se no limiar previamente citado. Ao identificar os possíveis itens que irão compor a lista de opções incorretas, é verificado se existem itens suficientes. Se houver, esses serão utilizados para a geração do exercício. Se não, são buscados termos secundários que são sinônimos do sinônimo trabalhado.

5.2.5 API de Retorno do Exercício

Após a geração dos exercícios, eles são enviados para o usuário para que possa haver a interação. Cada exercício gerado é tratado como uma requisição REST. Os dados são comunicados entre o servidor e a interface de usuário (*user interface* - UI) por meio de dados JSON.

Para os exercícios gerados, é enviada a interface de usuário apenas uma *string* JSON que contém os seguintes itens: um número identificador, a pergunta e a lista de opções geradas. Para ocorrer a verificação da resposta correta, a UI deve realizar uma consulta informando o identificador da pergunta e a lista de opções selecionadas pelo usuário. Por fim, o *backend* encaminha, para cada opção selecionada pelo usuário, se ela está correta ou incorreta. Desse modo, o *backend* da aplicação está finalizado, podendo o *frontend* se conectar a ele.

Capítulo 6

Resultados Experimentais

Neste capítulo, são apresentados os resultados experimentais das principais etapas da metodologia de extração de conhecimento e geração de exercícios. Para as análises, como uma referência de textos em língua inglesa, foi utilizada a base de dados *The Signal Media One-Million News Articles Dataset* (Corney et al.; 2016). Essa é uma base de dados de notícias, disponibilizada pela *Signal Media*[1] de modo a facilitar a realização de pesquisas sobre notícias. A base de dados contém um conjunto de aproximadamente 1 milhão de artigos jornalísticos que foram coletados no mês de setembro de 2015. A fonte dos dados é composta por jornais e blogs, em sua maioria escritos em língua inglesa (Corney et al.; 2016). Foram utilizadas 93.000 fontes únicas, como semente, para a coleta e destas foram coletados 265.512 artigos de *blogs* e 734.488 artigos de jornais. Cada artigo conta, em média, com 405 termos. Essa base de dados foi utilizada como fonte de termos a serem analisados.

Para realizar a análise sobre a metodologia desenvolvida, tornou-se necessário também a utilização de um dicionário da língua inglesa. O dicionário utilizado foi a *WordNet*, por ser uma base léxica amplamente utilizada em pesquisas de PLN e conforme justificado na Seção 2.2.4 da Metodologia.

Este capítulo é apresentado na seguinte estrutura: inicialmente, a Seção 6.1 apresenta a comparação entre três algoritmos de DLS. Em seguida, na Seção 6.2, são apresentados os resultados obtidos pelo Gerador de Sinônimos por Definição. Na Seção 6.3, são apresentados os resultados obtidos no processo de Seleção das Opções dos Exercícios. Por fim, na Seção 6.4, são apresentados os resultados obtidos pelo Gerador de Exercícios.

[1] http://signal.uk.com/

6.1 Algoritmos de Desambiguação de Sentidos

A metodologia do Gerador de Sinônimos por Definição necessita de um algoritmo de desambiguação de sentidos como parte integrante do Manipulador de Exemplos de definições, conforme ilustrado na Figura 5.1. Diversos trabalhos têm proposto algoritmos de desambiguação, dentre eles destacamos Li et al. (2012), Wang et al. (2014) e Banerjee and Pedersen (2002). Esses trabalhos foram selecionados por utilizarem abordagens diferentes para realizar a desambiguação de termos dentro de frases específicas, ou seja, em determinados contextos.

Para validar os algoritmos para a base utilizada nesse experimento, foram realizado dez experimentos na qual foram selecionadas 50 frases aleatórias da base de dados de frases *The Signal Media One-Million News Articles Dataset*. Para cada frase foi selecionada uma palavra, também aleatoriamente, para ser identificado o sentido utilizado na frase. A identificação da definição correta para cada termo foi realizada por um especialista. Ele foi responsável por ler todas as definições de cada termo e identificar aquela que é utilizada em cada contexto. Posteriormente, o algoritmo é executado e é realizada a conferência do resultado a partir das indicações realizadas pelo especialista, obtendo, assim, a acurácia do algoritmo. Por fim, a partir da média das dez execuções, analisou-se o melhor algoritmo para a essa base.

O primeiro algoritmo avaliado foi proposto por Banerjee and Pedersen (2002) e propõe uma abordagem que adapta o algoritmo *Lesk*, originalmente proposto por Lesk (1986). Esse algoritmo se baseia na suposição de que palavras em uma dada vizinhança tendem a dividir uma definição em comum. Basicamente, o algoritmo compara a frase a ser analisada com as frases de exemplo do dicionário. A versão adaptada utiliza conhecimentos do *WordNet* para melhorar a qualidade do algoritmo.

O segundo algoritmo avaliado foi proposto por Wang et al. (2014), o qual sugere uma abordagem supervisionada utilizando *semantic diffusion kernel*. Essa técnica consiste na criação de um modelo semântico, por similaridade dos significados. O modelo é implementado em grafo definido pela co-ocorrência dos termos e significados, além de informações léxicas. Busca-se, assim, obter um modelo semelhante ao *bag of words*, com grande representatividade de termos e exemplos, contudo, com relações mais suaves. Essa suavidade é obtida pelo uso de outras características da frase, além do conjunto de palavras, o que possibilita a ocorrência de um modelo mais completo.

O algoritmo de Li et al. (2012) também foi analisado no processo de seleção do

Proposta	Acurácia
Banerjee and Pedersen (2002)	85,00%
Wang et al. (2014)	80,00%
Li et al. (2012)	73,00%

Tabela 6.1: Comparação dos Algoritmos de Desambiguação.

desambiguador. A abordagem é diferente das anteriores por ser uma abordagem não-supervisionada que visa diminuir o custo computacional, utilizando modelos probabilísticos para a definição inicial dos parâmetros. Cada um dos três algoritmos utiliza técnicas diferentes para realizar a desambiguação lexical de sentido. A Tabela 6.1 apresenta o resultado da execução dos três algoritmos. Foi selecionado o modelo proposto por Banerjee and Pedersen (2002), por apresentar a melhor acurácia para os testes realizados.

6.2 Experimentos do Gerador de Sinônimos por Definição

O Gerador de Sinônimos por Definição é o componente da metodologia responsável por identificar relações de sinonímia entre dois termos. Essa relação é mais complexa do que uma lista de sinônimos, pois busca constatar a sinonímia de dois termos por meio da semelhança apresentada entre as definições. Logo, o Gerador de Sinônimos por Definição apresenta uma lista de sinônimos para o termo, identificando qual o sinônimo tem maior relação de acordo com cada definição do termo. Nesses experimentos, a métrica utilizada é a acurácia, uma vez que é esperado que o valor obtido experimentalmente seja refletido na base de conhecimento.

A primeira tarefa a ser realizada nesse componente consiste em identificar qual a definição do termo é utilizada em determinada frase. Para isso é utilizado o algoritmo de DLS presente no Manipulador de Exemplos de definições, conforme descrito na Seção 5.1.3.

Os experimentos do Gerador de Sinônimos por Definição iniciam na Seção 6.2.1, onde são apresentadas as configurações dos experimentos de sinonímia por definição. Em seguida, a Seção 6.2.2 apresenta as análises do Gerador de Sinônimos por Definição.

6.2.1 Configuração dos Experimentos de Sinonímia por Definição

Uma vez definido o desambiguador de definição a ser utilizado, iniciou-se o processo de analisar o Gerador de Sinônimos por Definição. Inicialmente deve ser realizada a configuração dos experimentos de sinonímia por definição. Essa configuração é realizada para selecionar a amostra a ser utilizada nos experimentos e rotular as definições, de modo a existir uma fonte de comparação às escolhas da metodologia proposta.

A amostra utilizada contém 50 termos que foram selecionados aleatoriamente da Base *The Signal Media One-Million News Articles Dataset*. As definições dos termos foi obtida da Base de Conhecimento. Cada termo contém uma lista de definições e para a realização do experimento foi selecionada uma definição para cada termo. A escolha da definição também foi aleatória. A Tabela 6.2 apresenta a distribuição do número de definições por termo da base de dados. Por exemplo, a base possui um termo que tem nove definições e três termos que possuem oito definições. Na tabela, é possível observar que poucos termos têm as maiores quantidades de definições, uma vez que as cinco maiores listas estão associadas a sete termos diferentes. Em contrapartida as menores listas estão associadas a um maior número de termos. Constata-se esse fato ao observar que as cinco menores listas estão associadas a 29 termos.

Selecionados os termos a serem utilizados, foi realizada a rotulagem deles. Essa rotulagem consiste em utilizar um especialista em língua inglesa para realizar o relacionamento entre o termo e um de seus sinônimos, selecionado aleatoriamente. O especialista é uma pessoa com conhecimento em língua inglesa e vivência de um ano na Inglaterra. O relacionamento consiste em analisar todas as definições do sinônimo e ponderá-las de acordo com a sua similaridade com a definição do termo sob análise. Foram utilizados pesos entre 0 (zero) e 3 (três), com definição a saber:

- 0: não há similaridade entre as definições;
- 1: similaridade baixa;
- 2: similaridade média;
- 3: similaridade alta.

Os pesos foram associados para cada definição do sinônimo. Não existe a obrigatoriedade de utilização de todos os pesos, entre 0 e 3, na lista de determinado termo.

# Definições	# Termos
20	1
19	1
18	1
17	1
16	3
14	1
12	3
11	2
10	1
9	1
8	3
7	1
6	3
5	5
4	5
3	8
2	8

Tabela 6.2: Distribuição do número de definições por termo em análise.

Rótulo	# Definições	# Termos
0	232	4
1	57	8
2	56	23
3	15	15

Tabela 6.3: Distribuição do número de definições e do número de termos para cada rótulo.

Logo, poderão existir casos em que não há correspondência entre as definições do termo analisado e dos sinônimos, obtendo-se, assim, uma lista na qual o peso associado a todas as definições é 0 (zero). Portanto, quando não houver definições que sejam amplamente semelhantes, o peso 3 (três) não é utilizado, ficando as definições com peso 2 (dois) como aquelas que melhor indicam a sinonímia para aquela definição. O especialista é responsável por realizar essa ponderação que será utilizada como referência para comparação com o resultado obtido pelo algoritmo.

Observando a Tabela 6.3, é possível visualizar que apenas 15 definições foram rotuladas com valor 3, enquanto 232 delas com 0. A distribuição dos rótulos pelos termos também é apresentada na Tabela 6.3. Nessa ótica, é possível observar que apenas quatro termos não apresentam um sinônimo. Mesmo que fraco, ou seja, rotulado pelo especialista com valor 1. Também apresenta que 15 termos têm sinônimos com similaridade máxima e 23 com similaridade nível 2.

6.2.2 Gerador de Sinônimos por Definição

A primeira fase da metodologia proposta, conforme descrito na Seção 5.1.3, consiste em gerar uma base de dados que relacione os termos da língua inglesa. O processo consiste em utilizar a frase de exemplo do sinônimo, substituindo o termo sinônimo pelo termo sob análise. Em seguida, executar o desambiguador. O resultado obtido foi a lista de definições do termo juntamente com um peso associado. A partir da lista definida pelo especialista e a lista resultante do algoritmo, foram comparados os resultados e, assim, obtidos os erros e acertos do algoritmo comparado com a definição do especialista. Inicialmente, buscou-se identificar se há uma relação dos pesos associados às definições e o acerto do algoritmo.

As análises foram divididas em três cenários distintos. Cada um dos cenários visa analisar os dados sob uma perspectiva de flexibilidade, do algoritmo, diferente. O primeiro consiste em analisar os casos nos quais o algoritmo selecionou a melhor opção de acordo com a indicação do especialista. O segundo cenário considera que as rotulagens com peso 2 e 3 não são consideradas como erro. Ou seja, para um termo que o especialista rotulou duas definições distintas com pesos, 2 e 3, respectivamente e o algoritmo escolheu aquela com peso 2 como resposta será também considerado com acerto. Por fim, o terceiro cenário considera que as rotulagens com peso 1, 2 e 3 não são consideradas como erro. Ou seja, para um termo que o especialista rotulou duas definições distintas com pesos 1 e 3, respectivamente, e o algoritmo escolheu aquela com peso 1 como resposta será também considerado com acerto.

Para essas análises, são considerados acertos as definições selecionadas pelo algoritmo, que vão de acordo com a escolha, prévia, do especialista. Por exemplo, no primeiro cenário, só é considerado acerto as escolhas realizadas pelo algoritmo para definições com rótulo 3, de acordo com o especialista. Em contrapartida, para o terceiro cenário, definições rotuladas com 1, 2 ou 3 são consideradas corretas quando indicadas pelo algoritmo. Ao contrário do acertos, os erros são as indicações do algoritmo que não estão de acordo com as indicações do especialista.

Em cada cenário são realizadas duas análises distintas. A primeira consiste em analisar os 50 termos selecionados aleatoriamente. No segundo cenário são analisados os resultados considerando 46 termos. Nesse são desconsiderados os termos que não têm sinônimos, ou seja, aqueles que todas as definições foram rotuladas com valor 0 pelo especialista. A Tabela 6.4 apresenta os dados a serem analisados em todos os cenários. Nela, a coluna "Rótulo" consiste nos valores utilizados para rotular os termos. Na coluna "Quantidade de termos" é possível visualizar a quantidade de termos que foram rotulados, pelo especialista, com cada valor. A coluna "Melhor resposta" indica a quantidade de termos nos quais a definição escolhida pelo algoritmo contém o maior peso utilizado pelo especialista na análise daquele termo. Ou seja, dos 24 termos que foram rotulados com 2, 16 correspondem à melhor resposta indicada pelo especialista. Com isso, para esses 16 termos, não houve nenhuma definição cujo rótulo é maior que 2. Por fim, a coluna "Erros reais" é uma subtração entre os valores contidos nas colunas "Quantidade" e "Melhor resposta", indicando a quantidade de termos nos quais o algoritmo selecionou uma definição cujo peso não representa a melhor definição para aquele termo, segundo o especialista.

Rótulo	Quantidade de termos	**Melhor resposta**	**Erros**
3	7	7	0
2	24	16	8
1	11	8	3
0	8	4	4

Tabela 6.4: Quantidade de indicações de melhor resposta, resposta errada por quantidade de termo e peso.

Para a primeira análise do cenário 1, os 50 termos são analisados e apenas o maior rótulo é considerado correto. Neste caso houveram 19 erros, obtidos através da soma dos valores da coluna "Erros" para os rótulos 1 e 2, somado a "Quantidade de Termos" rotulados com 0. Logo, houveram 35 indicações correta, obtendo, assim, a acurácia de 62,0%. Na segunda análise do cenário 1, na qual são desconsiderados os 4 termos que não houveram sinônimos identificados pelo especialista, logo, são consideradas as 46 instâncias com sinônimos, o número de erros é diminuído a 15. Esse valor é obtido pela soma dos valores exibidos na coluna "Erro". Com isso, são totalizadas 31 indicações corretas. Neste cenário a acurácia obtida é de 67,4%.

No caso da primeira análise do cenário 2, cujos rótulos de valor 2 e 3 são considerados corretos para 50 termos analisados. Somando os 3 erros do rótulo 1 às 8 rotulagens de valor 0, obtém-se o total de 11 erros e, consequentemente, 39 acertos do algoritmo. Neste caso, é atingida a acurácia de 78,0%. Para a segunda análise do cenário 2, no qual são considerados 46 termos, é calculado o total de 7 erros, obtido através da soma da coluna "Erros" das rotulagens com valor 0 e 1. Portanto o número de acertos é de 39 o que gera a acurácia de 84,8%.

Por fim, hão as análises do cenário 3, na qual os valores 1, 2 e 3, do rótulo, são considerados corretos. A primeira análise, considerando os 50 termos, apresenta um total de 8 erros representados pelo rótulo 0. Com isso, foram 42 os acertos do algoritmo, que representam uma acurácia de 84,0%. Na segunda análise, considerando 46 termos, o total de erros é diminuído a 4 e o número de acertos mantém-se em 42. Para este caso, a acurácia é de 91,3%.

Juntamente com a indicação da relação de sinonímia, o algoritmo de desambiguação, apresentado na Seção 5.1.3, atribui um peso que varia entre 0 e 1, indicando o quanto cada definição do sinônimo se aproxima da definição sob análise. Esses dados também

	Média	Mínimo	Máximo	Variância	Desvio Padrão
Acerto	0,1361	0,0205	0,4297	0,0096	0,0980
Erro	0,1196	0,0292	0,3313	0,0068	0,0824

Tabela 6.5: Análise do peso atribuído pelo algoritmo para as definições indicadas como corretas.

foram analisados. As Tabelas 6.5 até 6.8 demonstram os resultados da análise dos valores máximo, mínimo, de média, variância e desvio padrão dos pesos atribuídos pelo algoritmo. Esses dados são separados por acerto e erro do algoritmo. Os acertos consistem nas definições que o algoritmo atribuiu o maior peso em concordância com a atribuição realizada pelo especialista. Os erros representam a indicação oposta, na qual a indicação de melhor sinônimo realizada pelo algoritmo não condiz com aquela previamente realizada pelo especialista.

A Tabela 6.5 apresenta os dados das definições escolhidas pelo algoritmo como aquela de maior similaridade com o termo sob análise. Nessa tabela, é possível observar que o maior peso associado a uma definição, de 0,4297, representou uma indicação correta do algoritmo. Complementarmente, um valor também alto, de 0,3313, foi erradamente indicado como correto. O desvio padrão com valores próximos, com 0,0980 para os acertos e 0,0824 para os erros, indica que somente a partir do peso associado à melhor definição não é possível inferir se ela é a que mais se assemelha ao sinônimo. Os demais dados, com valores próximos para erros e acertos, confirmam essa informação. O desvio padrão indica que para os casos de acerto o peso atribuído pelo algoritmo estará mais distante da média do que nos casos de erro. Fato confirmado pela variância. Essas informações, unicamente, não fornecem embasamento suficiente para validar o algoritmo. De modo a fundamentar esses dados, novas análises foram realizadas.

De modo a validar se o peso associado à pior definição auxilia na sua validação, esse dado também foi analisado. A Tabela 6.6 demonstra os dados obtidos por essa análise. A média dos pesos associados a definições de termos nos quais o algoritmo errou foi de 0,0365, abaixo da média das listas nas quais o algoritmo acertou. Essa média indica que em alguns casos, a definição de menor semelhança ao sinônimo pode ter um valor próximo à definição de maior similaridade, demonstrando diversas definições semelhantes para o termo. A variância para os casos de erro é uma ordem de grandeza menor do que para os acertos, indicando que há uma menor variação entre os pesos dos casos de erro

	Média	Mínimo	Máximo	Variância	Desvio Padrão
Acerto	0,0441	0,0068	0,2353	0,0016	0,0397
Erro	0,0365	0,0127	0,1048	0,0008	0,0276

Tabela 6.6: Análise do peso atribuído pelo algoritmo para as definições com menor peso.

	Média	Mínimo	Máximo	Variância	Desvio Padrão
Acerto	0,0921	0,0092	0,3816	0,0075	0,0868
Erro	0,0831	0,0166	0,2265	0,0035	0,0593

Tabela 6.7: Análise da diferença de pesos entre a melhor e pior definição indicada pelo algoritmo.

do algoritmo.

O desvio padrão e a variância, dos casos de erro, indicam uma baixa variação do peso associado à média das demais definições. Esse fato indica que o peso médio associado à definição de menor similaridade é maior nos casos de erro do algoritmo. Isso infere que há uma proximidade entre os pesos das definições dos termos indicados erroneamente.

De modo a confirmar os dados observados acima, foi realizada uma análise da diferença entre o maior e menor pesos associados às definições. A Tabela 6.7 apresenta os resultados obtidos. É possível observar que a média dos pesos para os casos de acerto do algoritmo, de 0,0921, é maior do que para os casos de erro, de 0,0831. Este fato indica que o algoritmo tende a acertar quando a diferença entre os pesos é maior. Também é possível visualizar que o desvio padrão e a variância, para os casos de acerto, é maior do que para os casos de erro. Esses fatos indicam uma relação entre o aumento na diferença das definições com maior e menor peso associado, mas não geram um resultado conclusivo.

De modo a identificar se o número de definições associada ao termo se relaciona com o número de acertos do algoritmo, foi analisado o número de definições associadas aos acertos e erros do algoritmo. A Tabela 6.8 apresenta o resultado obtido. Nela, é possível observar que o número máximo de definições para os erros do algoritmo é 9. Os valores

	Média	Mínimo	Máximo	Variância	Desvio Padrão
Acerto	9,4737	2	20	29,7718	5,4564
Erro	4,5454	2	9	4,2722	2,0671

Tabela 6.8: Análise do número de definições dos termos.

da variância e do desvio padrão para os casos de erro são substancialmente menores do que para os casos de acerto, indicando que o algoritmo tende a errar mais para termos com até 9 definições. Logo, é observado que um maior número de definições indica a possibilidade maior de acerto, por parte do algoritmo. A partir dessa informação é possível inferir que um menor número de definições as torna mais genéricas, fato que tende a diminuir a acurácia do algoritmo ao buscar sinônimos analisando os sentidos.

A partir da análise dos dados apresentados, é possível observar que médias maiores dos pesos das definições tentem a gerar mais acertos por parte do desambiguador. Outro fator importante é o número médio de definições. Quanto mais definições, há uma maior diferença entre os pesos associados e, consequentemente, pode ser um facilitador na identificação das definições que correspondam àquela utilizada para cada contexto.

6.3 Experimentos de Seleção das Opções dos Exercícios

De modo a identificar as palavras que serão utilizadas para a geração de exercícios, foram analisadas as suas características. Esta análise visou gerar um algoritmo capaz de aumentar a precisão na identificação da resposta correta na geração de exercícios. As características analisadas foram: o número de definições do sentido indicado como correto, o total de frases de exemplo da palavra, o número de exemplos do sentido indicado como correto e a soma de todos os pesos associados ao termo. Para esta análise, a medida escolhida foi a precisão. Isso ocorreu porque espera-se que o algoritmo resultante maximize o número de termos candidatos para a geração de exercícios.

Inicialmente foi analisado o número de definições total que a palavra apresenta. Desse aspecto, obtemos os dados demonstrados no gráfico presente na Figura 6.1. Nele observamos que ao aumentar o número de definições, aumentamos também a precisão na escolha da resposta correta. Em contrapartida, o número de palavras que satisfaz a

condição é inversamente proporcional ao número de definições. Ou seja, ao incrementar o número de definições, há uma diminuição do número de palavras que satisfaçam esta condição. Para a Figura 6.1, inicialmente não há restrição do número de definições, indicado pelo número 2, logo, as 50 palavras podem ser utilizadas. Neste caso, a precisão obtida pelo algoritmo é de 78,0%.

Ao limitar o número mínimo de definições para 3, o número de palavras que atendem a condição é diminuído a 33 e, como consequência,a precisão, recebe um pequeno incremento atingindo 78,6%. Para 4 definições o número de palavras é reduzido a 27 gerando um novo aumento na precisão, o que resulta em um total de 79,4%. Definindo o número de definições como 6, o total de palavras reduz a 20, enquanto a precisão soma 87,0%. Limites entre 7 e 9 causa uma variação na precisão entre 95,0% e 93,8%, limitando o número de palavras a 15. A precisão de 100%, é atingida quando o número de definições é no mínimo 10, o que limitou o número de palavras em 14.

Neste ponto é possível concluir que o aumento no número de definições tende a aumentar a precisão na escolha dos termos. Em contrapartida, o número de termos a serem utilizados diminui. Logo, valores entre 7 e 9, são bons limitantes do número de definições, mantendo uma acurácia média de 94,5% e mantendo, em média, 35,0% das palavras.

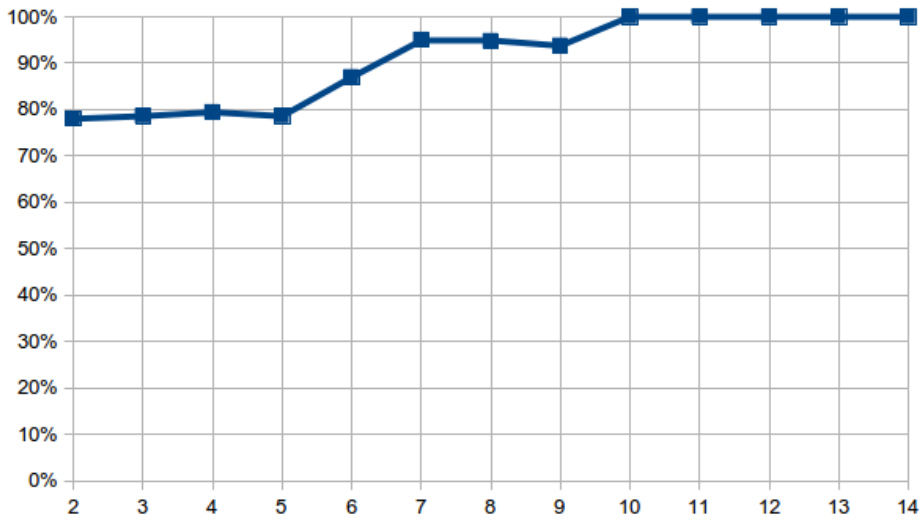

Figura 6.1: Acurácia do algoritmo X Número de definições do sentido indicado como correto.

A segunda característica analisada, consiste no número total de exemplos da palavra. O gráfico presente na Figura 6.2 apresenta o resultado dessa característica. Nele é

possível que nos casos em que o número de exemplos varia entre 2 e 10, a precisão alterna entre 77,1% e 80,8% e o número de palavras é limitado a 21. Para valores entre 14 e 18 exemplos, a precisão é incrementada, recebendo valores entre 93,8% e 94,7%. Contudo o número de palavras disponíveis cai a 16. Utilizando de um mínimo de 22 definições, a precisão atinge 100% para um uso de 22,0% das palavras.

O número total de exemplos da palavra demonstra o traço observado no número de definições total. O aumento na restrição da característica reflete com o aumento na precisão e diminuição no número de palavras disponíveis para uso.

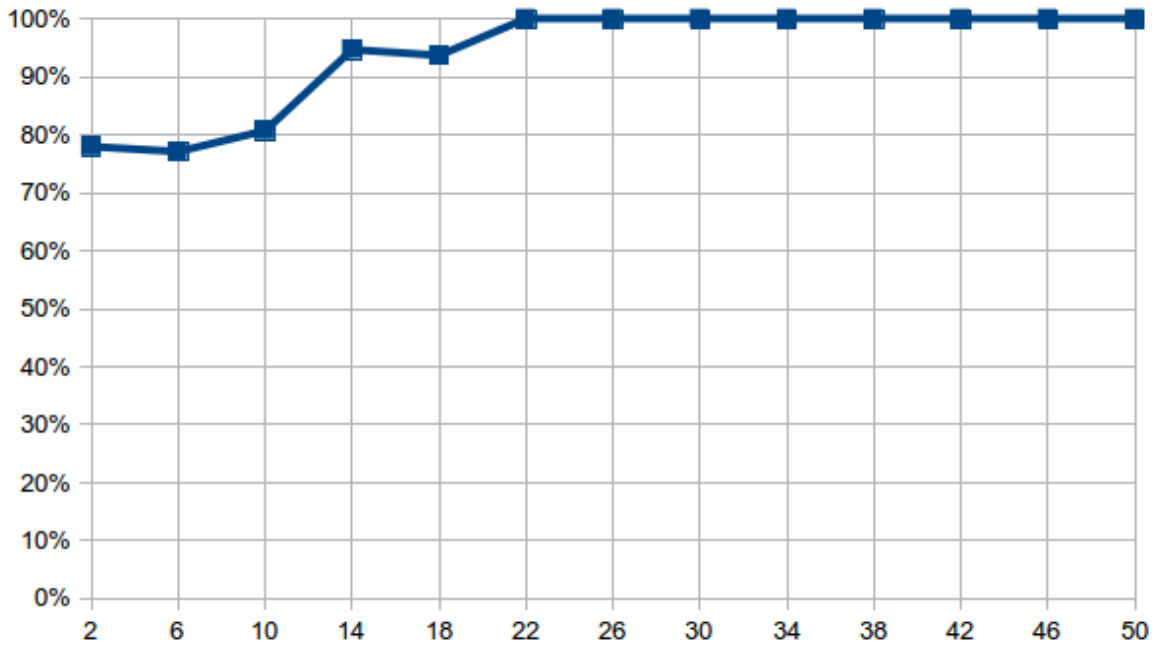

Figura 6.2: Acurácia do algoritmo X Total de exemplos da palavra.

O número de frases presente na definição indicada pelo algoritmo como correta foi outra característica analisada. O gráfico resultante é apresentado na Figura 6.3. Nele, é possível observar que a precisão sem limitação do número de definições é de 78,0%. Para os valores 2, 3 e 4 a precisão varia entre 89,5%, 85,7% e 100%, respectivamente. Baseando nos mesmos valores, o número de palavras a serem utilizados alterna entre 19, 7 e 6, respectivamente.

Essa análise demonstra valores um pouco inferiores às anteriores em relação à precisão, uma vez que a segunda melhor precisão atingida, de 89,5%, com uma taxa de uso de palavras de 38,0%. Ao considerar 3 exemplos, o número de palavras utilizada cai drasticamente, atingindo 7. Com isso, o uso dessa característica não demonstra ser

interessante como as demais analisadas.

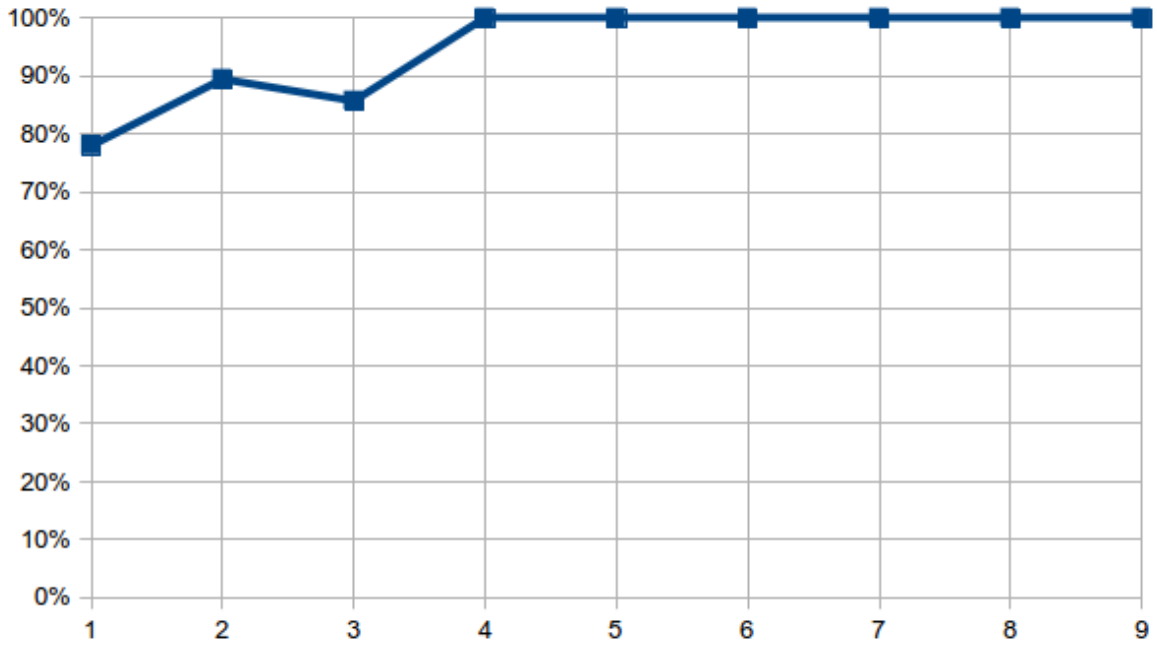

Figura 6.3: Acurácia do algoritmo X Total de exemplos do sentido indicado como correto.

Por fim, a quarta característica analisada consiste na soma dos pesos atribuídos pelo algoritmo às definições da palavra analisada. A Figura 6.4 apresenta o gráfico dessa análise. Nela é possível observar que o aumento no peso utilizado pelo algoritmo para distribuir entre as definições indica uma maior probabilidade de acerto. Esse fato é observado através da precisão crescente na figura, incrementando seu valor de 78,0% até 92,3%, com pequenas oscilações. Neste mesmo cenário, o número de palavras disponíveis para a geração de exercícios decrementa de 50 para 13.

Essa característica demonstra uma maior limitação no número de palavras para a geração de exercício, em conjunto com o número de frases presente na definição indicada pelo algoritmo. Isso é observado pelo fato de restringir a 15 palavras, totalizando peso 0,7, para atingir a precisão de 90,0%.

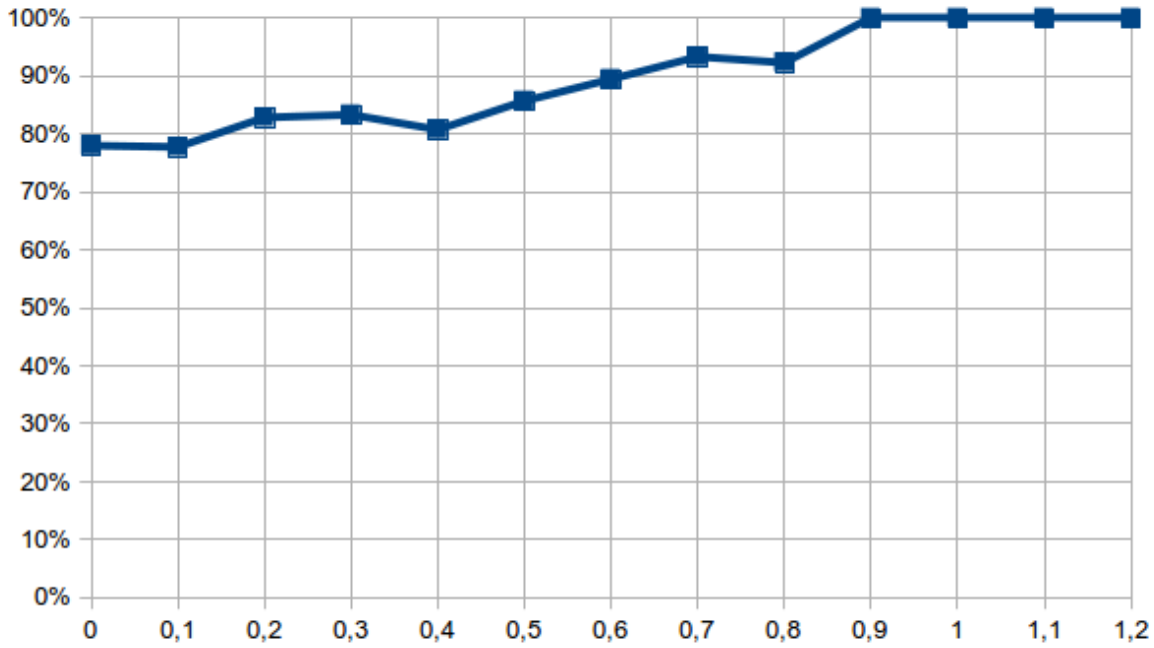

Figura 6.4: Acurácia do algoritmo X Soma dos pesos das definições.

6.4 Experimentos do Gerador de Exercícios

O Gerador de Exercícios visa utilizar a base de conhecimento gerada a partir do Gerador de Sinônimos por definição e automatizar a criação de exercícios referentes a vocabulário. Os experimentos realizados visam analisar a quantidade e a qualidade dos exercícios gerados pelo Gerador de Exercícios. Como quantidade, busca-se avaliar o número de exercícios gerados pelo Gerador de Exercícios. A qualidade condiz com uma maior quantidade de exercícios gerados corretamente, ou seja, nos quais as opções não sejam ambíguas e a opção definida como correta realmente seja a opção correta.

Esse experimento foi realizado utilizando 50 termos, diferentes daqueles utilizados na Seção 6.2, mas também obtidos da base *The Signal Media One-Million News Articles Dataset*, conforme supracitado.

Assim como no Gerador de Sinônimos por definição, busca-se identificar um fator que possibilite uma relação entre quantidade e qualidade satisfatória, de modo que o número de exercícios gerados seja abrangente o suficiente para testar o maior número de termos possível do texto. Também espera-se que os exercícios gerados não contenham indicações incorretas para o usuário.

Limiar	Número de Questões
0,0000	363
0,0010	262
0,0020	257
0,0025	257
0,0050	233
0,0100	232
0,0200	211
0,0250	189
0,0500	104
0,1000	47
0,1500	12
0,2000	5
0,2500	3

Tabela 6.9: Número de questões geradas por variação do limiar.

A Tabela 6.9 apresenta o número de questões geradas pelo Gerador de Exercícios, de acordo com a variação do limiar para seleção da opção correta. O limiar consiste na diferença do peso associado entre a primeira e a última definição para o termo. Ou seja, só é gerado o exercício de uma dada palavra quando a diferença entre os pesos da primeira e última definição, atribuído pelo algoritmo, é maior que o valor do limiar. Foram adotados valores entre 0,0 e 0,25, o que gerou entre 3 e 363 exercícios. Ainda na Tabela 6.9, é possível observar que o limiar obtido na Seção 6.2, de 0,05, gerou 104 exercícios. Também é possível observar que, para os limiares entre 0,0010 e 0,0100, há uma variação de 30 exercícios, enquanto que para limiares superiores a 0,0100 há uma queda mais acentuada do número de exercícios gerados. Realizando a comparação de ordens de grandeza, entre 0,0100 e 0,1000, há uma diminuição de 185 no número de exercícios gerados.

Os exercícios gerados foram divididos em três questões distintas. Cada questão gerou um número diferente de exercícios, conforme pode ser observado na Tabela 6.10. A questão 1 busca identificar qual dos significados listados nas opções representa a definição do termo destacado na frase presente na pergunta. A questão 2 insere definições de sinônimos dentre as opções e o usuário deve identificar qual a definição que pertence

Limiar	Questão 1	Questão 2	Questão 3
0,0000	59	171	90
0,0010	40	111	73
0,0020	35	111	76
0,0025	35	116	80
0,0050	37	98	70
0,0100	38	97	69
0,0200	25	93	66
0,0250	31	79	61
0,0500	12	46	39
0,1000	5	21	18
0,1500	3	5	4
0,2000	1	2	2
0,2500	1	1	1

Tabela 6.10: Número de questões gerada por exercício.

ao termo destacado. Por fim, a questão 3 insere uma lista de definições do termo e de seus sinônimos de modo que o usuário identifique aquelas que pertencem ao termo. Os tipos de exercícios gerados foram apresentados em detalhes na Seção 5.2.4. É possível observar que a questão 1 gerou menos exercícios que as demais para a maioria dos limiares utilizados. Isso ocorre porque foram utilizados apenas termos com mais de 5 sentidos.

Uma vez gerados os exercícios, foi analisado se as opções selecionadas tornam o exercício possível de resolução. Com isso, busca-se identificar se não há opções ambíguas dentre as listas de opções geradas. Essa condição foi analisada, por um especialista, para todos os exercícios gerados. A Figura 6.5 apresenta o gráfico com a variação do limiar e a porcentagem de exercícios gerados sem opções ambíguas. Nela, é possível observar que ao aumentar o limiar, o número de opções ambíguas é decrementado. A Figura 6.5 demonstra essa informação para cada questão gerada. Confrontado esses dados com aqueles apresentados na Tabela 6.10, também é observado que a qualidade dos exercícios é inversamente proporcional ao número de exercícios gerados. Ou seja, quanto menos exercícios gerados maior a qualidade do conjunto resultante.

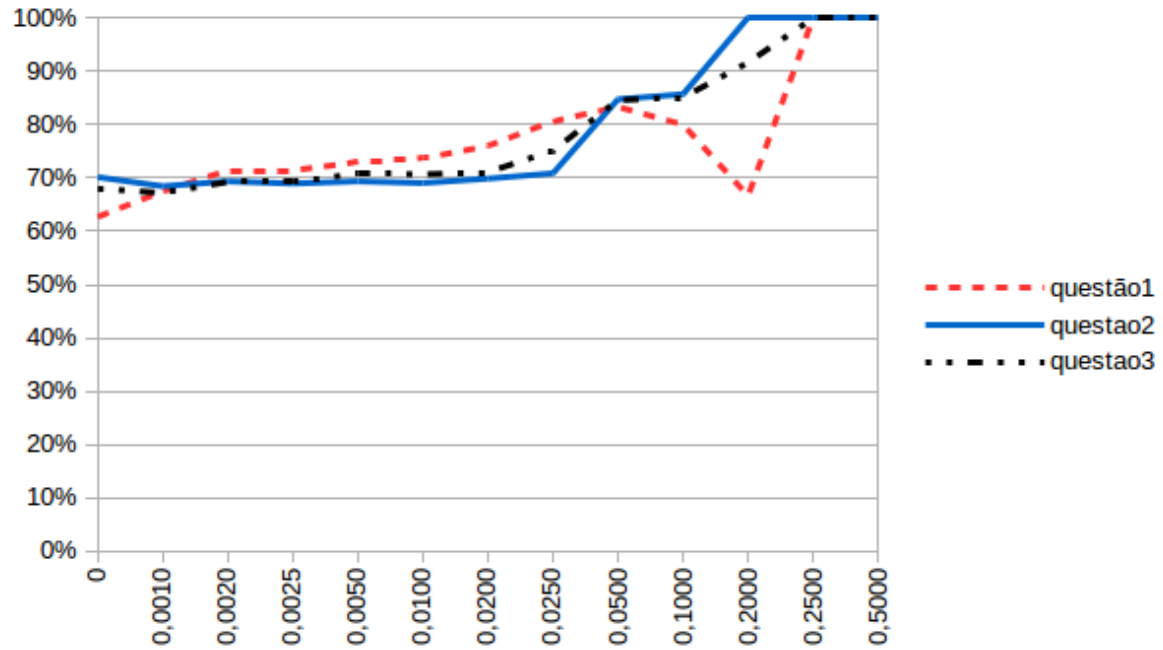

Figura 6.5: Acurácia por questão X variação do limiar de seleção da opção correta.

Ao agregar os dados presentes na Tabela 6.10, obtêm-se o gráfico presente na Figura **??**. Nela é apresentada a qualidade total dos exercícios gerados. Constata-se a partir do gráfico que aproximadamente 84,6% dos exercícios, com limiar de 0,0500, não contam com uma lista de opções ambígua. Isso corresponde a um total de 88 exercícios dos 104 gerados.

A partir dos dados analisados é observado que a partir do limiar 0,05, com uma acurácia de 84,6%, há o aumento dessa acurácia com a possibilidade de alcançar 100% ao utilizar o limiar de 0,2. Em contrapartida há uma diminuição do número de exercícios gerados, de 104 até diminuir a cinco. Nesse ponto, ao utilizar o sistema totalmente automatizado, é considerado ideal utilizar o limiar de 0,2, limitando o número de exercícios gerados de modo a manter a acurácia em seu valor máximo. Sob outra ótica, ao utilizar um sistema semi-automatizado, é ideal manter uma acurácia alta mas que possa ser validado por um utilizador humano. Nesse ponto, um limiar entre 0,05 e 0,2 provê uma acurácia superior a 84% com um número maior de exercícios que podem ser utilizados.

,,

,,,,

Capítulo 7

Conclusões e Trabalhos Futuros

Atualmente, os métodos para adquirir, ou aprimorar, conhecimentos em um novo idioma seguem técnicas tradicionais. Essas técnicas apresentam conteúdo engessado e com longos períodos de duração, o que pode causar desinteresse do estudante. Esses fatos indicam a necessidade de criação de um método de aprendizado com conteúdo diversificado, a fim de aumentar o interesse dos estudantes e, naturalmente, a sua capacidade de aprendizado, com maior absorção de conteúdos, já que são conteúdos de seu interesse. Para isso, este trabalho compreendeu o desenvolvimento de uma metodologia de geração de conhecimento, baseado em sinônimos, bem como a utilização desse conhecimento para a geração automática de exercícios.

Nessa visão, esse trabalho propôs o desenvolvimento de uma metodologia para ponderação de sinônimos de diferentes definições de um mesmo termo, visando a geração automática de exercícios de vocabulário de uma língua. Desse modo, torna-se possível uma opção automatizada de estudo de vocabulário de um novo idioma.

Para alcançar os objetivos, foi coletada a base de dados léxica *WordNet*, utilizada como dicionário e base de dados de sinonímia. A partir dessa base, foi gerada uma base de conhecimento na qual ocorre a ponderação dos termos sinônimos. As análises foram realizadas sobre a base de conhecimento sob três cenários distintos. A acurácia obtida variou entre 62,7% e 91,3%. Para o caso médio, a acurácia obtida foi de 78,0% ao analisar todos os termos selecionados e 84,0% quando foram desconsideradas as palavras sem sinônimos.

Como parte da geração de exercícios, foram analisadas algumas características das palavras a fim de identificar padrões que auxiliem na seleção de palavras. Os resultados

obtidos variaram entre 78,0% e 100% e todas as características demonstraram o seguinte padrão de comportamento: ao aumentar o limiar, a precisão é incrementada, enquanto o número de palavras disponíveis para a geração de exercícios diminui. Neste ponto, foi observado que utilizar o número total de definições que a palavra apresenta com valores entre 7 e 9, e o número total de exemplos da palavra entre 14 e 18, possibilita uma precisão entre 93,7% e 95,0% para um uso aproximado de 30% das palavras. A geração de exercícios também foi analisada em busca de identificar a sua acurácia. A acurácia resultante variou entre 68,0% e 100% e o número de questões alternou entre 363 e 5.

O trabalho demonstrou resultados consistentes na geração da base de conhecimento e na seleção das características utilizadas para selecionar as palavras a serem utilizadas na geração de exercícios. Uma vez que a acurácia aumenta, a medida em que se restringe os limiares de forma a limitar as possíveis palavras que serão utilizadas na geração de exercícios, a solução desenvolvida cumpre a função de geração de exercícios com uma qualidade desejada, como uma acurácia de 95,0%.

A inclusão de novos dicionários tende a aumentar a acurácia da solução proposta. Esse fato é corroborado pelos experimentos realizados, nos quais é possível observar que palavras com maior quantidade de definições correspondem a mais acertos por parte do algoritmo. Há também a possibilidade inclusão de outras características da palavra no processo de seleção para a geração de exercícios. Desse modo, limitar o número de palavras utilizadas no processo de geração de exercícios.

Como trabalhos futuros, para o Gerador da Base de Conhecimento, são sugeridos os seguintes pontos: (1) a utilização de outro dicionário como *thesaurus* de sinônimos, complementarmente ao *WordNet*, para aumentar a base de exemplos e assim aumentar também a acurácia; (2) implementação de outros algoritmos de desambiguação para substituir o atual, ou uma combinação dos algoritmos existentes; (3) realizar a análise para uma amostra maior; (4) gerar uma base de conhecimento baseada em antônimos, o que pode ser útil para a comunidade científica.

Com relação ao Gerador de Exercícios, são sugeridos como trabalhos futuros: (1) desenvolver de outra técnica de geração de exercícios; (2) adicionar outras características no componente de seleção de palavras para geração do exercício; (3) analisar um conjunto de amostra maior; (4) analisar a geração de exercícios com um conjunto diferente de limiares; (5) desenvolver uma aplicação real, que utilize da metodologia proposta; (6) contribuir com o *WordNet*, ao incluir pesos em todas as relações de sinonímia.

Referências Bibliográficas

Agirre, E. and Edmonds, P. (2007). Word sense disambiguation: Algorithms and applications.

Allan, J., Croft, B., Moffat, A. and Sanderson, M. (2012). Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne, *SIGIR Forum* **46**(1): 2–32.
URL: *http://doi.acm.org/10.1145/2215676.2215678*

An, J., Lee, S. and Lee, G. G. (2003). Automatic Acquisition of Named Entity Tagged Corpus from World Wide Web, *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2*, ACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 165–168.
URL: *http://dx.doi.org/10.3115/1075178.1075207*

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Baeza-Yates, R., Mayo-Casademont, M. and Rello, L. (2015). Feasibility of Word Difficulty Prediction, *String processing and Information Retrieval*, Springer International Publishing, pp. 362–373.

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet, *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, pp. 136–145.

Banerjee, S. and Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness, *Ijcai*, Vol. 3, pp. 805–810.

Bobrow, D. G., Fraser, J. and Quillian, M. (1967). Automated language processing, *Annual review of information science and technology* **2**: 161–186.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora, *Psychological Science* pp. 96–99.

Carr, E. M. and Mazur-Stewart, M. (1988). The effects of the vocabulary overview guide on vocabulary comprehension and retention, *Journal of Reading Behavior* **20**(1): 43–62.

Chowdhury, G. (2010). *Introduction to Modern Information Retrieval, Third Edition*, 3rd edn, Facet Publishing.

Chowdhury, G. G. (2003). Natural language processing, *Annual review of information science and technology* **37**(1): 51–89.

Clements, M., de Vries, A. P. and Reinders, M. J. (2008). Optimizing single term queries using a personalized Markov random walk over the social graph, *Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR) in ECIR'08*.

Conte, G., Casadei, D. I. B. and Nardi, N. L. (2012). Culture: A better manner to learn english, *Ágora: revista de divulgação científica* **16**(2esp.): 550–570.

Corney, D., Albakour, D., Martinez, M. and Moussa, S. (2016). What do a million news articles look like?, *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, pp. 42–47.
URL: *http://ceur-ws.org/Vol-1568/paper8.pdf*

Damerau, F. (1976). Automated language processing, *Annual review of information science and technology* **11**: 107–161.

Houaiss, A. D. E. H. d. (2003). da língua portuguesa, *Rio de janeiro, Objetiva* .

Jin, H. and Wong, K.-F. (2002). A Chinese Dictionary Construction Algorithm for Information Retrieval, *ACM Transactions on Asian Language Information Processing (TALIP)* **1**(4): 281–296.
URL: *http://doi.acm.org/10.1145/795458.795460*

Laufer, B. and Waldman, T. (2011). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English, *Language Learning* pp. 647–672.

Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification, *WordNet: An electronic lexical database* **49**(2): 265–283.

Lee, Y. K., Ng, H. T. and Chia, T. K. (2004). Supervised word sense disambiguation with support vector machines and multiple knowledge sources, *Senseval-3: third international workshop on the evaluation of systems for the semantic analysis of text*, pp. 137–140.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, *Proceedings of the 5th annual international conference on Systems documentation*, ACM, pp. 24–26.

Li, X., Zhao, X., Fan, F. and Liu, B. (2012). An improved unsupervised learning probabilistic model of word sense disambiguation, *Information and Communication Technologies (WICT), 2012 World Congress on*, IEEE, pp. 1071–1075.

Lin, C., Liu, D., Pang, W. and Wang, Z. (2015). Sherlock: A semi-automatic framework for quiz generation using a hybrid semantic similarity measure, *Cognitive computation* **7**(6): 667–679.

Manning, C. D., Raghavan, P., Schütze, H. et al. (2008). *Introduction to information retrieval*, Vol. 1, Cambridge university press Cambridge.

Martin, J. H. and Jurafsky, D. (2000). Speech and language processing, *International Edition* **710**.

Mihalcea, R., Corley, C. and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity, *AAAI*, Vol. 6, pp. 775–780.

Mihalcea, R. and Moldovan, D. I. (1999). A method for word sense disambiguation of unrestricted text, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, pp. 152–158.

Nguyen, K.-H. and Ock, C.-Y. (2013). Word sense disambiguation as a traveling salesman problem, *Artificial Intelligence Review* **40**(4): 405–427.

O'Neil, H. F. and Perez, R. S. (2013). *Web-based learning: Theory, research, and practice*, Routledge.

Plaza, L. and Diaz, A. (2011). Using semantic graphs and word sense disambiguation techniques to improve text summarization, *Procesamiento del lenguaje natural* **47**: 97–105.

Ramos, S. G. (2014). Métodos de ensino de segunda língua e língua estrangeira, *UNOPAR Científica Ciências Humanas e Educação* **1**(1).

Rello, L., Baeza-Yates, R., Bott, S. and Saggion, H. (2013). Simplify or Help?: Text Simplification Strategies for People with Dyslexia, *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, ACM, New York, NY, USA, pp. 15:1–15:10.

Sakaguchi, K., Arase, Y. and Komachi, M. (2013). Discriminative approach to fill-in-the-blank quiz generation for language learners., *ACL (2)*, pp. 238–242.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*, McGraw-Hill computer science series, McGraw-Hill, New York.
URL: *http://opac.inria.fr/record=b1091083*

Sardinha, T. B. (2004). *Lingüística de corpus*, Editora Manole Ltda.

Sildus, T. I. (2006). The effect of a student video project on vocabulary retention of first-year secondary school german students, *Foreign Language Annals* **39**(1): 54–70.

Sung, L.-C., Lin, Y.-C. and Chen, M. C. (2007). An automatic quiz generation system for english text, *Seventh Ieee International Conference On Advanced Learning Technologies (Icalt 2007)*, IEEE, pp. 196–197.

Tan, L. (2014). Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software], https://github.com/alvations/pywsd.

Wang, T., Rao, J. and Hu, Q. (2014). Supervised word sense disambiguation using semantic diffusion kernel, *Engineering Applications of Artificial Intelligence* **27**: 167–174.

Zuccon, G., Koopman, B. and Bruza, P. (2014). Exploiting Inference from Semantic Annotations for Information Retrieval: Reflections From Medical IR, *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '14, ACM, Shanghai, China, pp. 43–45.
URL: *http://doi.acm.org/10.1145/2663712.2666197*