UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DESCOBERTA AUTOMÁTICA DE EXPRESSÕES MULTIPALAVRAS A PARTIR DE TEXTOS PARALELOS

NATALIE LOURENÇO VARGAS

ORIENTADORA: PROFA. DRA. HELENA DE MEDEIROS CASELI
CO-ORIENTADOR: PROF. DR. CARLOS RAMISCH

São Carlos – SP

Agosto/2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

DESCOBERTA AUTOMÁTICA DE EXPRESSÕES MULTIPALAVRAS A PARTIR DE TEXTOS PARALELOS

NATALIE LOURENÇO VARGAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Inteligência Artificial

Orientadora: Profa. Dra. Helena de Medeiros Caseli

São Carlos – SP

Agosto/2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Natalie Lourenço Vargas, realizada em 11/10/2018:

Helena de M. Caseli

Profa. Dra. Helena de Medeiros Caseli
UFSCar

Profa. Dra. Vânia Paula de Almeida Neris
UFSCar

Profa. Dra. Aline Villavicencio
UFRGS

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Aline Villavicencio e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Helena de M. Caseli

Profa. Dra. Helena de Medeiros Caseli

A Deus e a todas as pessoas que me apoiaram até aqui.

AGRADECIMENTOS

Gostaria de agradecer o CNPq, que me concedeu bolsa de estudos nos primeiros meses desse mestrado, e à FAPESP, como membro da equipe dos projetos AIM-WEST (Auxílio Regular #2013/50757-0) e MMeaning (Auxílio Regular #2016/13002-0). Gostaria também de agradecer a empresa Accenture que apoiou-me na continuação do desenvolvimento deste trabalho. Agradeço também aos meus orientadores Helena e Carlos, que além de grandes professores e pesquisadores, são uma grande inspiração na minha carreira. Por último agradeço a Deus e a todas as pessoas que estiveram presente me dando forças nessa jornada.

Epígrafe

Yoda

Always pass on what you have learned.

RESUMO

Expressões Multipalavras (EMs) são um desafio atual para a área de Processamento de Linguagem Natural e existem diferentes métodos automáticos propostos para descobri-las e tratá-las nos textos. Neste trabalho propomos dois métodos de descoberta de EMs de forma bilíngue em textos paralelos, os quais foram implementados em uma ferramenta que recebeu o nome de *Bilingual Discovery MWE Toolkit* (BiDiMWEToolkit). Com embasamento em ideias similares na literatura, os métodos propostos utilizam vetores de palavras (*word embeddings*) bilíngues para encontrar as melhores traduções para as EMs descobertas automaticamente. No primeiro método proposto, as EMs fonte e alvo são extraídas separadamente, a partir de padrões morfossintáticos pré-definidos, e, em seguida, ocorre o paralelismo das candidatas com base nos vetores bilíngues de palavras. No segundo método, a extração de EMs ocorre apenas no lado fonte, seguida da definição das melhores traduções (EMs ou não) alvo com base nos vetores bilíngues de palavras. Nos experimentos apresentados neste trabalho, para o par de idiomas português-inglês, concluímos que os dois métodos são capazes de realizar a descoberta bilíngue. O segundo método, contudo, apresenta duas vantagens em relação ao primeiro: (1) é capaz de gerar traduções sem a necessidade de ter conhecimento prévio da língua alvo, e (2) é capaz de gerar traduções contendo apenas uma palavra, abrangendo os casos em que EMs não são traduzidas necessariamente como expressões.

Palavras-chave: Expressões Multipalavras, EM, descoberta bilíngue, descoberta monolíngue, córpus paralelo

ABSTRACT

Multiword Expressions (MWEs) are a current challenge for Natural Language Processing field and there are different proposed automatic methods to treat and discovery them. We propose in this work two new bilingual discover methods in parallel texts, which were implemented as the Bilingual Discovery MWE Toolkit (BiDiMWEToolkit). The proposed methods were based on similar ideas in related works and they use bilingual word embeddings in order to find the best MWEs translations automatically discovered. In the first method, source and target MWEs are extracted separately from morphossyntatic patterns already defined and they are paired based on billingual word embeddings. In the second method, we just extracted source MWEs and the best translations are defined using bilingual word embeddings. As a result of our presented experiments, we concluded that both methods are capable of performing billingual discovery but the second method has prove to be more complete than the first method: (1) it capable of generating translations without target MWEs, so it wasn't necessary to have prior knowledge about the target language, (2) and capable of generating translations composed by one word, covering the cases when MWE translations are not an expression.

Keywords: Multiword Expressions, MWE, bilingual discovery, monolingual discovery, parallel corpus

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Frequentemente usamos palavras em nosso dia a dia que sozinhas possuem um sentido específico, mas que quando combinadas criam um sentido completamente diferente. Por exemplo, quando dizemos que há um *olho mágico* na porta não significa que a porta criou olhos com poderes mágicos; ou então quando vamos *tirar foto* não significa que iremos retirar ou remover algo chamado foto. Além disso, nossas conversas estão cheias de expressões que são claras apenas para quem possui um conhecimento mais vasto de uma língua. Por exemplo, quando dizemos *chutar o balde*, significa que vamos desistir de algo ou vamos perder o controle, enquanto que um estrangeiro aprendendo nossa língua pensaria no ato literal da expressão primeiro. Por outro lado, se estivermos no processo de aprender inglês e nos depararmos com a expressão *Hold your horses!* poderíamos nos perguntar por que alguém estaria nos mandando segurar nossos cavalos! Porém, uma vez que já estejamos mais familiarizados com a língua saberemos que a expressão *Hold your horses!* é, na verdade, uma maneira de dizer acalme-se.

Essas combinações de palavras são chamadas de **Expressões Multipalavras** (EMs) ou *Multiword Expressions (MWE)*, em inglês. EMs constituem um grande desafio para a área de Processamento de Linguagem Natural (PLN), principalmente para aplicações como a tradução automática. Por exemplo, *take a shower* em inglês significa *tomar banho* em português, mas se traduzirmos a primeira palavra de forma literal, o resultado poderia ser *pegar um banho*, o que não soa muito natural para a nossa língua. O exemplo *chutar o balde* citado no português, no sentido literal, seria traduzido como *kick the bucket* no inglês. Entretanto a expressão *kick the bucket* no inglês tem o sentido de morrer e não de desistir. A tradução apropriada para o inglês então seria *to throw everything up in the air.*

Quando usamos tradutores automáticos como auxílio para o aprendizado de uma segunda língua, é fundamental que antes do processo de tradução a ferramenta faça uma detecção automática de EMs, pois traduções literais de expressões nem sempre expressam a mesma coisa

que as expressões originais. Isso ocorre pois EMs apresentam, no geral, grau de composicionalidade baixo (CALZOLARI et al., 2002). Composicionalidade é o princípio que se usa para determinar a semântica de uma EM a partir do significado de seus componentes isolados. Essa detecção automática ocorre, em linhas gerais, detectando-se quão frequente é essa expressão em textos e a forma como as palavras que a compõem ocorrem juntas e de forma separada. Para definirmos as chances dessas expressões identificadas pela ocorrência no texto serem consideradas de fato EMs, usamos medidas estatísticas que calculam as frequências de ocorrência das palavras de forma separada e junta, pois é necessário verificar se duas palavras aparecerem juntas caracteriza uma expressão ou é apenas uma combinação de palavras que ocorre frequentemente na língua. Por exemplo, é possível verificar em um texto a ocorrência frequente da sequência *por uma*, mas será que a ocorrência dessas duas palavras caracteriza uma EM ou são apenas duas palavras aparecendo juntas muitas vezes?

Para deixar claro o que são essas construções que figuram como o objeto principal de estudo neste trabalho, a próxima subseção (1.1) traz definições e exemplos de EMs. A relevância das EMs para algumas aplicações de PLN é o tema da subseção 1.2, enquanto um relato breve e introdutório sobre como essas EMs são descobertas automaticamente nos textos é apresentado na subseção 1.3. Por fim, tem-se a apresentação do objetivo e da hipótese que norteiam este trabalho (subseção 1.4) e a organização do restante deste documento (subseção 1.5).

1.1 O que são EMs

Sag et al. (2002, p.2) definem EMs como "interpretações idiossincráticas que cruzam as fronteiras entre as palavras (ou espaços)". Já Calzolari et al. (2002, p. 1934) definem uma EM como "uma sequência de palavras que atua como uma única unidade em algum nível de análise linguística".

A dificuldade na descoberta automática de EMs reside no fato de que não há nenhum indício linguístico ou estrutural explícito para determinar se duas ou mais palavras juntas podem ser consideradas uma EM (SAG et al., 2002). Por exemplo, a aceitabilidade da expressão *aspirador de pó* ao invés de *aspirador de poeira* está relacionada apenas com seu uso frequente na língua, e não com a existência de alguma regra ou estrutura gramatical específica.

Intuitivamente, podemos identificar uma EM realizando a seguinte pergunta: "É possível trocar uma palavra da expressão por um sinônimo e manter seu sentido?". Por exemplo, a expressão *lua cheia* poderia ter seu sentido mantido se falássemos *lua completa*? Para um falante nativo de português, por exemplo, parece que não existe o mesmo sentido. Quando essa

troca acontece e o sentido se perde, existe o indício de que a expressão é, na verdade, uma EM.

Existem diferentes tipos de EMs em diferentes categorias. A EM *lua cheia* é um **composto nominal** uma vez que é uma combinação que se comporta como um substantivo. *Chutar o balde* por sua vez é uma **expressão idiomática** por ter um sentido completamente diferente da simples combinação dos sentidos das palavras que a compõem. *Ódio mortal* é uma **colocação**, pois, apesar do sentido da EM ser diretamente derivado do sentido de seus componentes, seus componentes são colocados lado a lado frequentemente na língua. Há também as **construções verbo suporte** como *fazer uma apresentação*, formadas por um verbo e um complemento nominal, onde o verbo não possui uma carga elevada de semântica e dá informações sobre tempo, pessoa e modo, e onde o complemento dá o sentido da ação, ou seja, o complemento dá o maior sentido da expressão como um todo.

Neste trabalho, o foco está nas construções verbo suporte, mas vale mencionar que a abordagem proposta pode ser aplicada/estendida para outros tipos de EMs.

1.2 EMs no Processamento de Linguagem Natural

O processamento de EMs em um texto é importante pois tem influência direta no desempenho de diversas tarefas de PLN, como:

1. Recuperação de Informação: Quando vamos fazer uma busca na internet por uma imagem de uma EM, pode ser que o buscador não entenda isso como uma EM e nos mande uma foto do significado literal ou imagens das palavras de forma segmentada. Por exemplo, se fizermos uma busca por imagens da palavra *mesa redonda* as primeiras imagens a serem mostradas serão de mesas com formato redondo e não de pessoas agrupadas para uma reunião. Acosta, Villavicencio e Moreira (2011) aplicam técnicas para melhorar a extração de EMs e demonstram, com base nos resultados de seus experimentos, que tratar EMs como uma unidade melhora a tarefa de recuperação de documentos.

2. Aprendizagem de uma Língua Estrangeira: EMs podem ser um problema para falantes não nativos de uma língua, pois sua memorização dificulta o aprendizado de uma língua nos seus diversos contextos de uso. Dicionários ou outros recursos lexicais que possuam entradas para EMs ajudam para que se possa aprender mais e evitar erros comuns. Por exemplo a palavra em inglês *six pack* pode ser utilizada para descrever aquele abdômen bem malhado e definido. Para o português, o equivalente seria *barriga tanquinho*. Traduzir *six pack* de forma literal não nos dá muitos indícios de que a palavra se refere ao

abdômen se não tivermos um conhecimento mais coloquial da língua.

3. Tradução Automática (TA): EMs representam uma grande dificuldade para sistemas de tradução automática uma vez que, frequentemente, traduzir algo palavra por palavra ocasiona um erro na tradução. O reconhecimento das EMs pelos sistemas de TA é importante para que a tradução da expressão seja efetuada como uma unidade, sem traduzir as palavras isoladamente. Além disso, o tratamento adequado de uma EM gera uma saída mais natural (fluente) na língua alvo (RAMISCH, 2012). Por exemplo, *pay a visit* no inglês significa *fazer uma visita* no português, mas a tradução literal e mais frequente da palavra *pay* é *pagar*, resultando na tradução incorreta (não equivalente) para *pagar uma visita*.

Além disso, linguistas têm realizado estudos com o objetivo de mostrar o quão presentes EMs estão em coleções de textos em diferentes línguas e domínios (BIBER et al., 1999). Também assume-se frequentemente que um léxico de um falante nativo possui tantas EMs quanto palavras simples (JACKENDOFF, 1997). Resultados empíricos e léxicos computacionais frequentemente confirmam essa hipótese, mostrando que EMs ocorrem frequentemente em vários registros de línguas, tanto na parte escrita quanto oral, em domínios gerais e específicos. Por exemplo, na WordNet[1] em inglês, dos seus 117.827 substantivos, 60.292 (51,4%) são EMs; e dos seus 11.558 verbos, 2.829 (25,5%) são EMs (RAMISCH; VILLAVICENCIO; KORDONI, 2013).

1.3 Descoberta automática de EMs

A descoberta de EMs de modo automático tem sido um tópico de interesse na comunidade de linguística computacional por bastante tempo (CHOUEKA, 1988; CHURCH; HANKS, 1990). Contudo, essa tarefa ainda permanece sem solução uma vez que é bastante complexa considerando-se que na maior parte do tempo não há indícios morfológicos e/ou sintáticos que indiquem a presença de uma EM em um texto. Por essa razão, medidas estatísticas são frequentemente empregadas para detectar ocorrências de tais expressões em um texto em uma língua. Por exemplo, é possível que o termo *descoberta* seja confundido com o termo *identificação* de EMs, por ser um assunto relativamente novo e suas definições na literatura estarem ainda sendo assimiladas por trabalhos que abordem esses temas (CONSTANT et al., 2017). Em geral a tarefa de descoberta consiste em encontrar novas candidatas a EM e gravá-las em uma lista para uso futuro, por exemplo um léxico. A tarefa de descoberta pode usar estratégias como medidas de associação e não usa nenhuma lista ou anotação de EMs conhecidas como base, que são justamente as estratégias adotadas neste trabalho. A identificação por sua vez, consiste em um

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

processo de anotação automática de EMs em um texto corrente, frequentemente necessitando um modelo aprendido a partir de um córpus de entrada anotado com EMs conhecidas. Zampieri et al. (2018) descrevem a construção de um sistema desenvolvido para identificar EMs do tipo verbal, usando córpus anotados com EMs do mesmo tipo, para até 19 idiomas.

As medidas de associação geralmente comparam essas ocorrências para avaliar qual a probabilidade dessas ocorrências serem apenas combinações aleatórias. Quando uma combinação de palavras é muito frequente, isso pode significar que ela não é apenas uma combinação aleatória. As medidas de associação dão valores altos para o caso extremo de uma combinação de palavras aparecer de forma frequente e seus componentes nunca aparecem sozinhos no texto, pois isso significa que a combinação é, sem sombra de dúvida, uma EM. Por exemplo, algumas medidas de associação possuem resultados que variam de 0 a 1. Se tivermos uma expressão cujos termos sempre aparecem juntos, e raramente de forma separada, a tendência é termos um valor convergindo para 1, indicando que essa expressão tem grandes chances de ser considerada uma EM. Por outro lado, quanto mais o valor da medida se aproxima de 0, menos chances uma combinação de palavras possui de ser considerada uma EM. Além disso, questões como o tamanho do córpus devem ser levadas em consideração na hora de usarmos tais medidas, pois esse fator pode influenciar o cálculo. Se o nosso córpus for muito pequeno, teremos poucas amostras de ocorrências de certos tipos de combinações, dificultando a análise das amostras.

Medidas estatísticas são calculadas com base em números de ocorrências estimados a partir de textos monolíngues. Quando a descoberta de EMs é realizada em textos de apenas uma língua, ela é chamada de "descoberta monolíngue". Entretanto, outras estratégias são empregadas para modelos de descoberta de EMs, dentre elas o uso de textos em uma segunda língua (SALEHI; COOK, 2013; TSVETKOV; WINTNER, 2014). Uma vez que textos correspondentes em duas ou mais línguas são usados para a descoberta de EMs, chamamos isso de "descoberta bilíngue ou multilíngue". Algumas pesquisas usam, além do alinhamento das sentenças de dois textos, o alinhamento de palavras para achar correspondências de tradução entre duas línguas (CASELI et al., 2010). Verificamos também o uso recente de vetores de palavras (*word embeddings*) para recuperar e analisar a semântica das expressões em um espaço vetorial (SALEHI; COOK; BALDWIN, 2015) ou usando dicionários para comparar as traduções literais das expressões com suas traduções oficiais, identificando que quanto menor é a similaridades entre traduções literais e as oficiais, maior é a probabilidade da expressão original ser uma EM (PEREIRA et al., 2014).

Para ilustrar o uso de textos em duas línguas na descoberta de EM, considere as sentenças em português retiradas do Córpus FAPESP com suas suas sentenças traduzidas para o inglês. Por exemplo, o item 1 corresponde à sentença em português e 1' corresponde a sua sentença

traduzida.

1. “Uma delas será capaz de **tirar fotos** e filmar em alta resolução”, diz Costa.

1’. “One of these will be able to **take** high resolution **photos** and videos,” says Costa.

2. “A idéia é **dar ênfase** às ciências, línguas, criatividade”, diz Nicolelis.

2’. “The idea is to **place emphasis** on the sciences, languages and creativity”, Nicolelis says.

3. “Tirapeli sabe que cada **foto** do seu livro conta histórias de objetos onde estão depositados o sagrado, a cultura, a história. ”

3’. “Tirapeli knows that each photo in his book tells stories of objects in which sacredness, culture and history have been deposited.”

4 Mas busca **tirar** lições dos êxitos e fracassos do programa.

4’ “However, one should learn from the program’s successes and failures.”

5. “Hoehne fez a maioria das **fotos**, mas há também trabalhos de outros naturalistas e fotógrafos.”

5’. “Hoehne took most of the **pictures**, but there are also photographs by other naturalists and photographers.”

6. “Novos meios como TV a cabo e a internet **tiraram** da novela o seu caráter de arena de problematização.”

6’. “New media, like cable TV and the Internet, have **taken away** from the soap opera its character as an arena for raising issues.”

Verificando os exemplos de número 1 e 2, as expressões *tirar fotos* e *dar ênfase* foram traduzidas também como expressões: *take photos* e *place emphasis*. Entretanto, verificamos que as palavras *tirar* e *fotos* podem assumir outros tipos de tradução como vemos nos casos de número 5 e 6, onde a palavra *tirar* é traduzida como uma expressão no inglês (*take away*) chamada de *phrasal verb*. Os casos acima mostram que assumir traduções literais das palavras que compõem uma EM pode gerar traduções incorretas da EM como um todo.

1.4 Objetivo e Hipótese

Este trabalho tem com objetivo propor um novo método para a descoberta bilíngue de expressões multipalavras em textos paralelos, com foco em construções verbo-suporte.

Para tanto, este trabalho se baseia na hipótese de que é possível recuperar o paralelismo entre EMs e suas traduções usando vetores bilíngues de palavras. O método proposto consiste em usar o paralelismo de textos alinhados por sentença, onde o primeiro texto está em português (fonte) e o segundo é a sua tradução em inglês (alvo). A ideia é verificar como as candidatas a EM em português estão traduzidas nas sentenças paralelas e usar essa correspondência para o método de descoberta bilíngue.

Apesar de alguns trabalhos usarem uma segunda língua para a descoberta automática de EMs, a saída dos sistemas propostos geralmente consiste em uma lista na primeira língua com suas respectivas pontuações, e raramente temos uma lista referente às possíveis traduções de cada EM descoberta em que cada par recebe uma pontuação que mostre as chances da candidata alvo ser a tradução da candidata fonte. Este trabalho, então, propõe-se a criar um método de descoberta automática de EMs usando uma segunda língua e também gerar uma lista de possíveis traduções para cada EM automaticamente identificada. Essa lista de traduções pode ser útil para a construção de um dicionário bilíngue, por exemplo. Além do uso do córpus bilíngue, usaremos os vetores de palavras e medidas estatísticas.

1.5 Organização

Este trabalho está organizado da seguinte maneira. O Capítulo 2 traz a fundamentação teórica necessária para o entendimento deste trabalho, com: uma descrição geral das abordagens propostas na literatura para a descoberta automática de EMs, assim como mecanismos a serem usados para este fim (vetores de palavras e alinhamento de palavras) e a forma de avaliação das EMs descobertas. O Capítulo 3 traz um apanhado geral dos trabalhos relacionados com este trabalho. O Capítulo 4 faz uma descrição da abordagem de descoberta automática de EMs proposta e sua arquitetura. Os experimentos para avaliar a abordagem proposta são apresentados no Capítulo 5. Por fim, no Capítulo 6 são feitas as considerações finais deste trabalho.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Como já mencionado anteriormente, EMs podem ser definidas como grupos de palavras que juntas possuem um significado diferente do que o obtido se seus significados fossem considerados isoladamente como, *chutar o balde*, *tomar banho* ou *mesa redonda*.

Existem diversas outras definições formais na literatura que explicam o que são EMs. Sag et al. (2002) definem uma EM como "interpretações idiossincráticas que cruzam os limites das palavras (ou espaços)". Baldwin e Kim (2010) afirmam que "EMs são itens lexicais que podem ser decompostos em lexemas e apresentam idiossincrasias sintáticas, semânticas, pragmáticas e/ou estatísticas." Para esse trabalho, vamos adotar a definição de Ramisch (2012) que é uma combinação das definições de Calzolari et al. (2002) e Baldwin e Kim (2010). Essa combinação define uma EM como itens lexicais que: 1) Podem ser decompostos em múltiplos lexemas; 2) apresentam um comportamento idiomático em algum nível da análise linguística e, como consequência, 3) devem ser tratados como uma unidade em algum nível do processamento computacional.

A seguir são apresentados alguns exemplos de categorias de EMs, existindo outros tipos também:

1. Compostos Nominais: quando as EMs se comportam como substantivos como em *olho mágico*, *full moon* (lua cheia) *aspirador de pó* e *casa branca*.

2. Expressões Idiomáticas: quando atuam com um significado específico pertencente a uma língua de tal maneira que se forem traduzidas literalmente para outra língua perdem seu significado. Muitas vezes as expressões idiomáticas estão relacionadas a gírias ou seu uso está restrito a determinados grupos sociais, de faixa etária etc. Por exemplo, a expressão em inglês *It's raining cats and dogs* na sua tradução literal significa *Está chovendo gatos e cães*, entretanto no português usamos *Está chovendo canivetes*.

3. Colocações: são combinações frequentes ou usuais de palavras, por exemplo: *ódio mortal* e *permanent job* (emprego fixo).

4. Nomes Próprios: formados por mais de uma palavra. Por exemplo, *Rio Grande do Sul* e *São Paulo*.

5. Construção Verbo Suporte (CVS): Segundo Savary et al. (2017) as CVS são formadas por um verbo e um substantivo, sendo que o último depende diretamente do verbo ou é introduzido por uma preposição. Por exemplo: *give a lecture* e *estar com fome*. Este substantivo também é um predicativo e frequentemente refere-se a um evento (ex: visita ou decisão) ou um estado (ex: coragem, alegria, medo). Exemplos: *make a decision*, *ter coragem*. O verbo que compõe a CVS é *leve*, o que quer dizer que contribui para o significado da EM como um todo apenas indicando o tempo e modo. Na expressão *fazer ginástica* o significado está concentrado em *ginástica*. Algo semelhante acontece com *dar um grito* na qual o significado da expressão está em *grito* e o verbo *dar* apenas oferece um suporte indicando que a ação (de gritar) foi realizada.

   No inglês, *Light Verb Construction* é a categoria que terá um foco especial nesse trabalho como demonstram os experimentos realizados até o momento. Um dos motivos para a escolha desse tipo de EM para investigação neste trabalho é a lacuna de trabalhos que tratam particularmente deste tipo de EM na literatura, pois quando fazemos buscas por trabalhos que tratem de EMs, a maioria deles tratam de EMs do tipo Colocação ou Expressões Idiomáticas.

2.1 Descoberta Automática de EMs

Os métodos propostos para a descoberta automática de EMs podem ser agrupados em dois grandes grupos: os monolíngues e os bilíngues. De modo geral, os métodos da abordagem monolíngue visam primeiramente identificar expressões candidatas com base em padrões morfossintáticos. Esses padrões morfossintáticos são definidos por meio de uma combinação das classes de palavras. Por exemplo, podemos fazer uma busca por candidatas a EM que sejam verbos seguidos de adjetivo, ou que sejam bigramas do tipo substantivo seguido de substantivo. Em seguida, esses métodos monolíngues realizam uma filtragem das expressões candidatas usando medidas de associação com o objetivo de eleger as expressões com mais chances de serem consideradas EMs. Os métodos da abordagem bilíngue, por sua vez, em geral usam a informação de uma possível tradução da expressão para realizar a filtragem. Alguns trabalhos utilizam corpora bilíngues (textos traduzidos em duas línguas diferentes) para realizar a extração. Alguns

trabalhos realizam o alinhamento de palavras, para identificar possíveis traduções no texto alvo, outros alinham apenas as linhas do texto (alinhamento sentencial) para realizar a descoberta. Mais recentemente, tanto os métodos da abordagem monolíngue quanto os da bilíngue têm usufruído dos vetores de palavras (SALEHI; COOK, 2013; GARCIA; GARCÍA-SALIDO; ALONSO-RAMOS, 2017) .

As próximas subseções trazem a fundamentação teórica por trás dos métodos de descoberta automática de expressões multipalavras. A seção 2.2 apresenta brevemente a abordagem tradicional monolíngue que se baseia em padrões morfossintáticos e medidas de associação. As técnicas usadas nas abordagens bilíngues, foco principal deste trabalho, são apresentadas em seguida: alinhamento de palavras (seção 2.3) e os vetores de palavras (seção 2.4). Por fim, a seção 2.5 traz as principais medidas usadas para avaliar os resultados desses métodos.

2.2 Abordagem Monolíngue

A abordagem tradicional para descoberta de EMs é a monolíngue, realizada em duas etapas: (i) extração de candidatas e (ii) filtragem das mesmas candidatas. Normalmente, a primeira etapa se dá através do uso de padrões morfossintáticos. Muitos trabalhos realizam um pré-processamento do córpus (por exemplo para etiquetação morfossintática) e, em seguida, aplicam os padrões para encontrar as candidatas. Na segunda etapa, as candidatas extraídas são filtradas com base em medidas de associação, que são fórmulas matemáticas capazes de capturar o grau de conexão ou associação entre os componentes de uma sequência de palavras (HOANG; KIM; KAN, 2009).

Padrões morfossintáticos são o método mais simples de extrair candidatas a EMs, uma vez que podem ser definidos por meio de expressões regulares. Para entender como os padrões morfossintáticos funcionam, considere o seguinte exemplo. Suponha que desejamos encontrar todas as ocorrências em que o verbo *realizar* seja uma construção verbo suporte, e por consequência, seja considerado uma EM. Para tanto, definimos o padrão *realizar* + SUBSTANTIVO. A partir desse padrão, o algoritmo então irá realizar uma busca por todo o córpus (previamente etiquetado morfossintaticamente) para recuperar todas as combinações que correspondam a esse padrão.

Além da etiquetação morfossintática para determinar a categoria gramatical das palavras (SUBSTANTIVO, VERBO, etc.), os textos muitas vezes passam por um processo de lematização das palavras para facilitar a descoberta. A lematização consiste no processo de desflexionar uma palavra para determinar sua forma canônica. Por exemplo, podemos encontrar o lema *realizar*

*Nos anos seguintes, Crick **realizou pesquisas** em hidrodinâmica e, durante a Segunda Guerra Mundial, projetou minas acústicas e magnéticas no Laboratório de Pesquisas do Almirantado, da Marinha Real Britânica.*
*Em 1992, Ysao Yamamura criou na Unifesp o Setor de Medicina Chinesa e Acupuntura, que além de **realizar pesquisas** atende os casos de dores ósseas e musculares agudas.*
*Além disso, a Fundação algumas vezes **realizou reuniões** com pesquisadores, líderes de projeto etc., e eu pretendo fazer isso mais sistematicamente.*
*Uma delas trata da criação de uma primeira empresa de propósito específico (EPE) pela Embrapa, que permitirá a essa empresa pública associar-se com parceiros privados para **realizar pesquisas**.*
*Após **realizar pesquisas** no Centro de Controle de Venenos e Antivenenos da Organização Mundial da Saúde, em Liverpool, Inglaterra, em 1986, a pesquisadora deu início a estudos para a fabricação de uma vacina antiofídica a partir do veneno dessa serpente comum no Brasil.*

Figura 2.1: Exemplo de um conjunto de excertos de textos para descoberta de EMs. (AZIZ; SPECIA, 2011)

em ocorrências como *realizei* e *realizaste*, entre outras flexões, tendo como lema *realizar*. O processo de lematização torna a busca de ocorrências com padrões morfossintáticos mais simples e eficiente pois recupera todas as ocorrências em que o lema é *realizar* .

Após a busca retornar todas as combinações correspondentes, as mesmas sofrerão um processo de cálculo que irá dizer se cada candidata tem chances altas ou baixas de ser uma EM. Por exemplo, se recebermos *realizar pesquisa* e *realizar armário*, iremos calcular quais as chances delas terem aparecido juntas por serem uma EM ou de ocorrerem apenas por uma combinação aleatória. EMs também ocorrem no texto de forma descontígua, quando seus componentes estão separados por outras palavras de outras classes gramaticais, como por exemplo *realizei uma reunião* tendo o artigo *uma* entre os componentes *realizei* e *reunião*.

Um outro exemplo de padrão morfossinático seria o que busca por compostos nominais do tipo SUBSTANTIVO + ADJETIVO no córpus. Esse padrão retornará candidatas do mesmo padrão que *olho mágico* por exemplo. Nesse caso, todas as candidatas recuperadas também passariam por uma filtragem indicando as chances de serem EMs de fato. O objetivo dessa filtragem é remover combinações de palavras regulares, como *olho esquerdo*, que correspondem ao padrão morfossintático de extração mas não são uma EM pois são completamente regulares, composicionais e previsíveis.

Para ilustrar esse processo, tomaremos como exemplo um conjunto de excertos de textos formados com frases extraídas do córpus bilíngue FAPESP (AZIZ; SPECIA, 2011) apresentado na Tabela 2.1. As cinco sentenças em português recuperadas do corpus FAPESP possuem ocorrências do padrão *realizar* + SUBSTANTIVO, destacadas em negrito.

Após a extração das candidatas a EM usando os padrões morfossintáticos, o próximo passo, então, é usar as medidas de associação para definir quais candidatas têm mais chances de serem consideradas, de fato, EMs. As medidas de associação testam a independência da ocorrência das palavras, ou seja, considerando um bigrama candidato a EM na forma p_1p_2, a palavra p_1 e a palavra p_2 serão vistas como duas variáveis aleatórias discretas. Em seguida, os cálculos estatísticos tentarão mostrar se as duas palavras (p_1 e p_2) são independentes e, como tal, podem coocorrer juntas em situações em que seja apenas coincidência, ou se na verdade elas são dependentes, indicando que juntas elas formam uma EM.

Para que tais medidas sejam calculadas, algumas outras informações devem ser levadas em conta tais como:

- frequência $f(p)$ – número de vezes que uma palavra ou sequência de palavras p aparece no córpus;
- tamanho T – tamanho do corpus (número total de *tokens* no córpus);
- probabilidade $P(p)$ – probabilidade de uma palavra ou sequência de palavras p estimada pela frequência relativa de p com relação ao tamanho T do córpus.

$$P(p) = \frac{f(p)}{T} \tag{2.1}$$

As frequências observadas para combinações de bigramas p_1 e p_2 são organizadas em uma **tabela de contingência** (PEARSON, 1904). Em uma tabela de contingência, as frequências podem ser entendidas com apresentado na tabela 2.1.

Tabela 2.1: Tabela de contingência

	p_2	$\neg p_2$	Somas
$p1$	$f(p_1, p_2)$	$f(p_1, \neg p_2)$	$f(p_1)$
$\neg p_1$	$f(\neg p_1, p_2)$	$f(\neg p_1, \neg p_2)$	$f(\neg p_1)$
Somas	$f(p_2)$	$f(\neg p_2)$	$f(T)$

A interpretação das informações nessa tabela é a seguinte: $f(p_1, p_2)$ é o número de vezes que o bigrama p_1p_2 ocorre no texto; $f(p_1, \neg p_2)$ indica o número de vezes que p_1 ocorre mas não está acompanhada de p_2 à direita; enquanto $f(\neg p_1, p_2)$ é o número de vezes que p_2 ocorre à direita, mas não está acompanhada de p_1 à esquerda; e assim por diante.

Com base nessa tabela, diversas medidas de associação são definidas como as apresentadas nas próximas subseções. Veja que as medidas são apresentadas considerando-se bigramas, ou

seja, EMs compostas por duas palavras. A ideia pode ser estendida para EMs de tamanhos maiores (trigramas, etc.).

2.2.1 Pointwise Mutual Information

Esta medida analisa o grau de dependência de duas variáveis aleatórias. A *Pointwise Mutual Information* é a razão entre a probabilidade de observação de uma palavra p_2 acompanhada de uma palavra p_1, e a probabilidade de observação de p_2 e p_1 separadamente (CHURCH; HANKS, 1990). A equação 2.2 apresenta esse cálculo, no qual $p1$ e $p2$ são as palavras que formam um bigrama candidato a EM.

$$PMI(p_1, p_2) = log_2 \frac{P(p_1, p_2)}{P(p_1)P(p_2)} \tag{2.2}$$

Na fórmula acima, as probabilidades do bigrama p_1, p_2 e de cada palavra p_1 e p_2 são estimadas de acordo com suas frequências relativas em um corpus, conforme definido na equação 2.1.

Vamos considerar o texto da Figura 2.1 como exemplo para entendermos como essa medida funciona. O primeiro passo é extrair uma candidata, e para isso, elegemos a expressão *realizar pesquisa*. Essa candidata é uma construção verbo suporte e a nossa extração levará em conta todas as formas em que a expressão ou as palavras que a compõem apareceram no texto, não apenas as suas formas base. Além disso, levamos em conta o fato de que nem sempre uma ocorrência é contígua, isto é, poderiam aparecer elementos no meio dessa expressão e ela continuaria válida, como em *realizou um grande número de pesquisas*.

Para fins de cálculo consideremos a palavra *realizar* como p_1 e a palavra *pesquisa* como p_2. Além disso, considerando o número total de *tokens* nas cinco sentenças desse mini corpus temos T=160. Assim, as probabilidades são:

$P(p_1, p_2) = \frac{4}{160} = 0,025$

$P(p_1) = \frac{5}{160} = 0,03125$

$P(p_2) = \frac{4}{160} = 0,025$

$PMI(p_1, p_2) = log_2 \frac{0,025}{0,03125 * 0,025} = 5$

PMI tem seu valor máximo quando duas palavras ocorrem apenas juntas no texto (nunca separadas). Além disso, essa medida não tem valor mínimo uma vez que $log0$ é indefinido. Para trigramas ou mais a fórmula pode ser reaplicada para probabilidades das três ou mais palavras

juntas dividida pelas probabilidades individuais.

É importante deixar claro aqui que, para todas as medidas de associação discutidas nesse trabalho (essa e as restantes), quanto maior o valor, mais chances uma candidata tem de ser considerada uma EM.

2.2.2 Log Likelihood Ratio

Essa medida foi proposta por Wilks (1938) e trabalha com a tabela de contingência (Tabela 2.1) na tentativa de medir a diferença entre a distribuição de probabilidade dos valores observados e dos valores esperados. A equação 2.4 mostra como esse cálculo é feito na qual os valores f_{ij} observados podem ser obtidos diretamente da tabela de contingência (2.1):

$f_{11} = f(p_1, p_2)$

$f_{12} = f(p_1, \neg p_2)$

$f_{21} = f(\neg p_1, p_2)$

$f_{22} = f(\neg p_1, \neg p_2)$

Já a frequência esperada é computada como se cada dado fosse independente, a parte $f_{ij}^{expected}$ da equação 2.4 pode ser definida assim:

$$f_{ij}^{expected} = \frac{f(dado_i)f(dado_j)}{T} \tag{2.3}$$

onde $dado_i$ e $dado_j$ podem ser p_1, p_2, $\neg p_1$ ou $\neg p_2$, dependendo de qual pedaço da tabela estamos usando. A medida *log likelihood ratio* é então definida pela fórmula abaixo, para todas as células $i \in \{1,2\}$ e $j \in \{1,2\}$ da tabela de contingência.

$$LL = 2\sum_{ij} f_{ij} log \frac{f_{ij}}{f_{ij}^{expected}} \tag{2.4}$$

Para calcular o valor de LL no texto de exemplo, vamos aplicar primeiro os valores antes do somatório:

$f_{11} * log \frac{f_{11}}{f_{11}^{expected}} = 4 * log \frac{4}{\frac{5*4}{160}} = 6,0205$

$f_{12} * log \frac{f_{12}}{f_{12}^{expected}} = 1 * log \frac{1}{\frac{5*156}{160}} = -0,6879$

$f_{21} * log \frac{f_{21}}{f_{21}^{expected}} = 0 * log \frac{0}{\frac{155*4}{160}} = 0$

$f_{22} * log \frac{f_{22}}{f_{22}^{expected}} = 155 * log \frac{155}{\frac{155*156}{160}} = 1,7043$

Fazendo então o somatório dos resultados e multiplicando por 2 temos, então, que o resultado do LL será 14,0738.

A faixa de resultados dessa medida será de algum valor maior que zero até um valor grande (sempre positivo). Quanto maior esse valor, melhor.

Para trigramas, essa medida acrescentaria a terceira palavra, ao invés da fórmula contar com ij, acrescentaríamos o k, usando um cubo de contingência ao invés de uma tabela de 2 dimensões. A fórmula seria baseada nas ocorrências $f_{(ijk)}$, e o denominador das frequências esperadas seria ajustado para T^2 ao invés de T..

2.2.3 Coeficiente Dice

Esse coeficiente mede se duas palavras não ocorrem juntas ao acaso, uma vez que esse índice seja alto, a candidata será uma boa candidata para EM. Essa medida é muito similar ao PMI, porém usa as frequências ao invés do logaritmo das probabilidades.

$$Dice = \frac{2 * f(p_1, p_2)}{f(p_1) + f(p_2)} \tag{2.5}$$

No caso do exemplo da Figure 2.1, o valor do coeficiente de Dice seria:

$Dice = \frac{2*f(p_1,p_2)}{f(p_1)+f(p_2)} = \frac{2*4}{4+5} = 0,8888$

Supondo que uma candidata apareça apenas uma vez no córpus e seus elementos apareçam apenas uma vez também, o valor máximo será 1. De maneira geral, poderia-se afirmar que, se uma candidata aparece x vezes e seus elementos nunca aparecem sozinhos no texto, então o valor do coeficiente de Dice será igual a 1, dando um indício de que realmente estamos vendo uma EM. Uma vez que temos estabelecido esse valor máximo, a faixa de valores será de 0 até 1 e quanto mais próximo o valor de uma candidata de 1, mais chances de ser uma EM.

Para os casos do trigrama ou n-gramas de tamanho arbitrário usamos, ao invés do valor 2, o valor n. Ao invés de calcularmos a frequência conjunta e separada de dois valores, aplicamos para três ou mais.

2.2.4 Student's T-score

Outra medida de associação bastante utilizada é a *t-score*, que estima os casos em que a co-ocorrência de um grupo de palavras é mais marcante em comparação com a sua co-ocorrência casual aleatória, de acordo com o valor da estatística de um teste de hipóteses cuja hipótese nula H_0 a rejeitar supõe que ambos seriam iguais. Computa-se, então, a diferença entre a frequência observada com o valor estimado de um bigrama dividida pela raiz quadrada da frequência observada.

$$t - score = \frac{f(p_1, p_2) - f_{11}^{expected}}{\sqrt{f(p_1, p_2)}} \tag{2.6}$$

Para o exemplo da Tabela 2.1, o valor do *t-score* seria:

$t - score = \frac{4 - \frac{5*4}{160}}{\sqrt{4}} = 1,9375$

Essa medida também pode ser adaptada para n-gramas e sua faixa pode variar de valores muito baixos (negativos) para valores muito grandes (na faixa positiva). Quanto maior seu valor, mais chances uma candidata terá de ser EM.

2.3 Abordagem bilíngue: alinhamento de palavras

Os métodos de descoberta automática de EMs da abordagem bilíngue baseiam-se na ideia de extrair padrões considerando, também, um segundo texto como uma fonte de informação complementar. Esse segundo texto pode ser tanto um córpus paralelo como comparável em outro idioma.

De maneira geral, trabalhos como o de Tsvetkov e Wintner (2012) usam o alinhamento de palavras para realizar comparações entre dois córpus de línguas distintas. O alinhamento de palavras consiste na tarefa de identificar as possíveis traduções entre palavras em um córpus bilíngue. O resultado pode ser visualizado como um grafo com as arestas ligando os nós que representam as palavras nos dois lados de um par de textos em dois idiomas. Por exemplo, considerando um trecho do mesmo texto apresentado no começo da seção, um alinhamento com uma tradução possível são apresentados na figura 2.2.

Observando o alinhamento da figura 2.2, verificamos que o termo *os* não tinha correspondência do lado alvo, sendo alinhada com o NULL (vazio). Alguns termos do lado alvo não foram alinhados também.

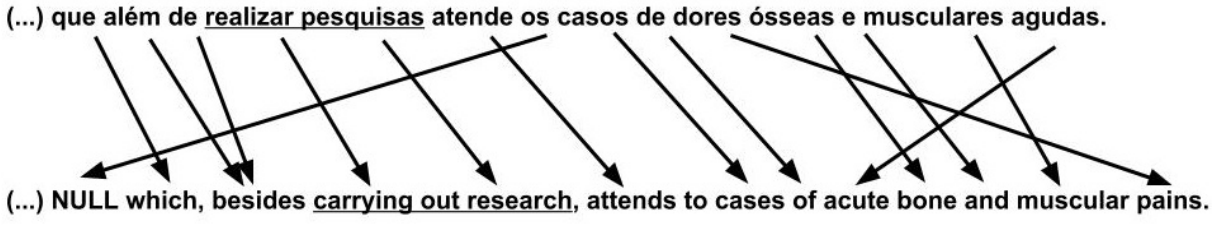


Figura 2.2: Exemplo de sentença paralela em português e inglês.

O alinhamento de palavras é tipicamente feito em um corpus paralelo alinhado por sentenças. Uma das estratégias mais usadas para se obter o alinhamento de palavras é por meio dos modelos estatísticos da IBM (BROWN et al., 1993) implementados na ferramenta GIZA++ (OCH; NEY, 2003). Como este trabalho não usou essa ferramenta, detalhes do seu funcionamento podem ser encontrados no *website* do GIZA++[1].

Nos experimentos apresentados neste documento, nós não usamos até o momento alinhadores de palavras mas queremos deixar claro que não somos contra a ideia de incluir essa informação no futuro. O motivo pelo qual escolhemos não incluí-lo até o momento é porque gostaríamos de verificar como a extração se comporta sem alinhadores, pois o resultado dos mesmos nem sempre é confiável.

Alguns trabalhos como Attia et al. (2010) não usam córpus paralelo para a extração de EMs, mas sim títulos curtos de páginas da Wikipedia em vários idiomas. Eles consideram que se uma página cujo título contém um equivalente para outra língua cujo título é apenas uma palavra, então o título da página original é provavelmente uma EM.

Em trabalhos como o de Pereira et al. (2014), dicionários bilíngues são usados para identificar possíveis traduções de candidatas de um lado fonte. Esses dicionários bilíngues são capazes de gerar traduções de forma automática (legíveis para máquina). Essas traduções são tomadas como recursos extras para determinar as chances de candidatas serem de fato EMs. É importante mencionar que cada palavra pode conter mais de uma tradução, então é necessário fazer combinações das mesmas, fazendo com que a tradução automática seja uma abordagem complexa. Comparando os valores da medida *Dice* para as candidatas alvo e fonte, a diferença de valores entre ambos foi um indício da natureza das candidatas fonte. Esse trabalho comparou essa abordagem com uma abordagem na qual só os valores de *Dice* do lado fonte eram considerados. O trabalhou usou o valor do *mean average precision* para comparar os dois métodos: o método monolíngue pontuou 0,54 enquanto que o método bilíngue pontuou 0,71. Esse resultado é um dos indícios dos benefícios da adoção de uma segunda língua para os métodos de descoberta de EMs.

[1] http://www.statmt.org/moses/giza/GIZA++.html

2.4 Vetores de Palavras

Vetores de palavras, também chamados de *word embeddings*, são representações distribucionais de palavras em um espaço real de *D* dimensões. Esses vetores são capazes de capturar a relação semântica entre palavras e têm sido usados em várias aplicações de PLN. Entre elas estão a identificação de várias relações morfossintáticas e semânticas (MIKOLOV et al., 2013), análise de sentimentos (SOCHER et al., 2013) e tradução automática (ZOU et al., 2013).

Representações distribucionais também têm sido usadas no contexto de descoberta de EMs. Salehi, Cook e Baldwin (2015) e Korkontzelos e Manandhar (2009) predizem o grau de composicionalidade de EMs através de vetores de palavras. Tais trabalhos usam vetores de palavras para medir a composicionalidade baseados na similaridade entre a expressão e os seus componentes no espaço vetorial. Para a criação desses vetores, a estratégia adotada é a de guardar a informação contextual presente no corpus em um vetor com dimensão pequena, sendo que cada palavra é representada por um vetor D-dimensional. . Essa ideia é bastante antiga, estando presente em modelos como Latent Semantic Analysis, propostos no início dos anos 90. No entanto, ela foi reintroduzida por Bengio (BENGIO et al., 2003), e ganhou popularidade com word2vec (MIKOLOV et al., 2013) depois com GloVe (PENNINGTON; SOCHER; MANNING, 2014). Em linhas gerais, os vetores de palavras criados com *word2vec* são modelos preditivos, uma vez que usam uma rede de neurônios para predizer uma palavra dado o seu contexto, representando-a de acordo com seus vizinhos.

Basicamente a estratégia para a geração desses vetores de palavras consiste em:

- Inicialmente é gerado um vetor aleatório para cada palavra no vocabulário;
- Em seguida, o método processa um córpus grande e, em cada passo, observa uma palavra e os seus contextos (palavras vizinhas dentro de uma janela);
- Palavras que estão em mesmo contexto tem sua similaridade maximizada;
- Após várias atualizações, os vetores que estão mais próximos uns dos outros mostram significados parecidos.

Word2vec: Consiste em um modelo de rede neural de duas camadas treinadas para reconstruir os contextos de adjacência das palavras. Proposto por Mikolov et al. (2013), a camada intermediária da rede neural é treinada para predizer uma palavra dado ao seu contexto (*Continuous Bag of Words*) ou um contexto dado uma palavra (*Skip-Gram*).

Quando uma dada palavra passa pela rede neural, a camada intermediária gera uma projeção desta palavra em um espaço vetorial. Como sua posição, nesta representação, está relacionada com seu contexto, palavras com significados similares terão distância vetorial menor.

O treinamento da rede neural é feito da seguinte forma. Inicialmente, cada palavra de um texto é traduzida em um vetor que terá o tamanho do vocabulário no córpus onde está inserida. Esse vetor será representado por V e contém zero em todas as casas, exceto na casa que representa a palavra alvo, cujo valor é igual a um (representação *one-hot*).

O modelo *CBoW* prediz o vetor V de uma palavra usando as palavras de seu contexto com o mesmo tipo de representação. A entrada será composta pela concatenação dos $N-1$ vetores contextuais e o objetivo será obter um vetor de tamanho V com a probabilidade de cada palavra a ser escolhida. Essa probabilidade irá representar a distribuição estatística do termo central. Já no Skip-Gram, cada palavra tenta predizer o seu próprio contexto. Assim, a entrada será o vetor da palavra em questão e a saída serão vários vetores de probabilidade, representando a distribuição estatística de cada palavra ocorrer próxima à palavra dada.

Por exemplo, considerando a candidata usada como exemplo no texto da Figura 2.1, recuperamos as palavras apresentadas na tabela 2.2 como as mais similares aos vetores gerados nos experimentos apresentados neste documento em um contexto monolíngue.

Tabela 2.2: Palavras alvo mais similares a uma palavra fonte de acordo com os vetores de palavras monolíngues

Palavra fonte	**Palavras alvo mais similares**
realizar	conduzir, executar
pesquisa	estudo, investigação
carry out	conduct, execute
research	survey, studies

Modelos de vetores de palavras também são capazes de gerar representações distribucionais usando palavras de duas línguas distintas. Córpus bilíngues são usados para que o modelo possa aprender e mapear em um espaço a relação semântica entre palavras de idiomas diferentes. O modelo irá se basear no número de vezes em que uma determinada palavra do lado fonte aparece no mesmo contexto do lado alvo. A tabela 2.3 apresenta os vetores geradas em um contexto bilíngue.

Escolhemos este método como recurso para a nossa descoberta bilíngue de EMs pois trabalhos têm adotado vetores bilíngues para geração de traduções de EMs. No trabalho de Garcia, García-Salido e Alonso-Ramos (2017), vetores de palavras bilíngues são usados para realizar a extração de colocações procurando o significado equivalente de uma língua para a outra. Esse

último trabalho é bastante similar ao nosso, como explicaremos no capítulo 3. Para ilustrar a questão da composicionalidade de EMs, abordada no Capítulo 1, vejamos o seguinte exemplo. A expressão *tomar banho*, já citada, tende a ser composicional, pois o substantivo *banho* tem o mesmo sentido que a expressão inteira. Se assumíssemos a composicionalidade como a única premissa para classificar uma EM, essa expressão não seria tomada como uma unidade (RAMISCH, 2012). Se traduzíssemos essa expressão, sem tomá-la como unidade, palavra por palavra, o resultado seria uma tradução literal diferente da expressão original (*take shower*).

As construções verbo-suporte, que são o foco deste trabalho, têm características composicionais mas ao serem traduzidas palavra por palavra, geram uma tradução final nem sempre correspondente a sua tradução oficial como expressão. Entretanto, um córpus paralelo é composto de um texto alvo que é a tradução feita por uma pessoa de um texto fonte, onde é possível que vetores bilíngues mapeiem traduções frequentes de palavras usando esse mesmo córpus alvo, e essas traduções não sejam necessariamente traduções literais retornadas por um dicionário, por exemplo. O uso de vetores de palavras são, portanto, recursos importantes para detectar o quanto construções verbo-suporte podem sofrer traduções de forma separada e o quanto isso pode impactar na descoberta bilíngue das mesmas.

Tabela 2.3: Palavras alvo mais similares a uma palavra fonte de acordo com os vetores de palavras bilíngues

Palavra fonte	Palavras alvo mais similares
realizar	carry out, conduct, execute
pesquisa	research, survey, studies
carry out	realizar, conduzir, executar
research	pesquisa, estudo, investigação

2.5 Avaliação

A avaliação da descoberta automática de EMs é uma tarefa difícil uma vez que a definição de uma EM depende de vários fatores, como vimos anteriormente. Existem algumas formas de se avaliar EMs, porém seu uso depende do contexto ou do que queremos tirar dessa avaliação. Assim, as próximas subseções trazem considerações sobre o contexto de avaliação (2.5.1) e o processo de anotação (2.5.2), finalizando com uma breve explicação das medidas comumente utilizadas (2.5.3).

2.5.1 Contexto de Avaliação

Os métodos de descoberta automática de EMs geram uma lista de candidatas a EMs. A avaliação dessas candidatas depende, então, do contexto, dos objetivos e dos recursos disponíveis. Alguns métodos de avaliação disponíveis foram escolhidos e retirados de Ramisch (2012):

Avaliação intrínseca : Os resultados são mostrados avaliando as EMs por si mesmas, diretamente como o produto final de um processo. Por exemplo, anotação das top-*k* candidatas (*k* candidatas melhor ranqueadas de acordo com alguma medida de associação) ou uso de um *gold standards* para calcular precisão e revocação de forma automática.

Avaliação quantitativa: Essa avaliação é feita usando medidas objetivas como precisão, revocação, medida F e *mean average precision*. Dessas medidas, precisão e *mean average precision* serão apresentadas na seção 2.5.3, uma vez que estas serão usadas na avaliação quantitativa neste trabalho.

Anotação manual: Um grupo de falantes nativos ou especialistas na língua irão decidir quais candidatas na lista gerada são, de fato, EMs. Nesse trabalho, usamos esse tipo de avaliação pois não há recursos disponíveis para anotação automática que corresponde a um outro tipo de anotação. Para obter maiores informações sobre anotação automática, consulte Ramisch (2012).

Avaliação baseada em *tokens*: Esse tipo de avaliação é realizada em qualquer caso que a EM é ambígua. Exemplos são os *phrasal verbs* do inglês e expressões idiomáticas. Nesse caso é necessário que um anotador analise o contexto daquela candidata para decidir se é ou não EM. Por exemplo, usando a expressão idiomática *chutar o balde* novamente, um anotador não pode inferir se essa candidata é ou não uma EM imediatamente pois isso depende de seu contexto. Isso significa que talvez essa candidata seja a ação literal de chutar o balde e não o significado metafórico. Existe também a avaliação baseada em categorias, na qual as candidatas são anotadas sem contexto pois as mesmas não apresentam ambiguidade. Para este trabalho os pares de candidatas a EM foram avaliadas com seus contextos originais do texto paralelo a fim de dar certeza e evitar ambiguidades no contexto de avaliar as candidatas e suas traduções.

2.5.2 Anotação

A anotação é uma parte muito importante da avaliação pois é ela que irá gerar o valor final de acertos do método de descoberta automática de EMs. A anotação manual, como mencionado na seção anterior, é feita por especialistas ou falantes nativos de uma língua. É necessário que haja pelo menos dois anotadores pois nem sempre o entendimento do que é ou não uma EM é o mesmo para todos os anotadores. Por isso a anotação final passará por uma análise de *concordância*, ou seja, os resultados finais da anotação passarão por um processo onde verifica-se o quanto as anotações estão parecidas.

A concordância entre anotadores é tradicionalmente calculada por meio do coeficiente *kappa* (*Fleiss' Kappa agreement score*) (FLEISS, 1971). Esse coeficiente gera um valor de concordância entre as instâncias anotadas em categorias fixas por dois ou mais anotadores. O coeficiente mede o quanto os anotadores concordaram em cada instância anotada e retorna a razão deste valor com o valor de concordância total esperada. Entretanto, esse coeficiente pode ser considerado um pouco controverso pois é necessário verificar o contexto da anotação para podermos decidir se um valor de $k \geq$0,6 é bom ou ruim.

2.5.3 Medidas de Avaliação

Para aplicarmos as medidas de avaliação, devemos olhar para a lista de saída das EMs que foram recuperadas pela descoberta automática de EM. Por exemplo, vamos tomar uma lista de candidatas com 100 candidatas (que serão chamadas de instâncias positivas) com suas entradas ordenadas pela medida T-score (do maior ao menor valor). Essa lista será chamada de *Lista C* e dentro dela vamos tomar as primeiras k candidatas denominadas $(C_{[1-k]})$. Para continuar com a ilustração, das 100 candidatas classificadas, vamos considerar que o k será de valor 80, ou seja, vamos tomar apenas 80 dessas candidatas para avaliação. O próximo passo então é considerar qual a taxa de EMs dentro dessa lista anotadas como Verdadeiras Positivas (VP, ou seja, anotadas como EMs). Essa taxa então é chamada de *Precisão em k* e é calculada da seguinte maneira:

$$P\ em\ k(C,k) = \frac{|VP\ em\ C_{[1-k]}|}{k} \tag{2.7}$$

Quando colocamos um valor pequeno, como 100 ou 200 para k, a anotação realizada por falantes nativos torna-se mais rápida. Mas nesse caso, a ideia é que se calcule a real precisão $P(C)$ anotando todo o conjunto dos candidatas retornadas, ao invés de pegarmos apenas $(C_{[1-k]})$. A precisão então vai ser a proporção de candidatas que foram julgadas como verdadeiras EMs

(PV) no conjunto de candidatas retornadas (C).

$$P(C) = \frac{|VP\ em\ C|}{C} \tag{2.8}$$

É possível ordenar essa lista de candidatas pela pontuação *T-score* verificando se as candidatas marcadas como verdadeiras EM estão no topo da lista. Realizando esse procedimento é possível aplicar a medida *Mean Average Precision* (MAP) a fim de analisar se as candidatas consideradas como EM têm de fato as melhores pontuações, verificando a qualidade da descoberta de EMs. Essa medida é útil pelo fato de que quando ordenamos uma lista de candidatas por alguma pontuação (T-score por exemplo), é difícil escolhermos um limiar das pontuações. Por exemplo, ordenando uma lista de 100 candidatas é difícil verificar até qual posição da lista estão as candidatas consideradas EMs. Com o MAP é possível verificar a qualidade da descoberta uma vez que ela calcula as vezes em que as candidatas consideradas EM estão mais ao topo da lista ordenada. A fórmula para o cálculo do MAP é a seguinte:

$$MAP(C) = \frac{\sum_{|C|}^{r=1} P\ em\ k(C,r) \times ehVP(C,r)}{|VPs\ em\ C|} \tag{2.9}$$

Onde a função $ehPV(C,r)$ se dá por :

$$ehPV(C,r) = \begin{cases} 1 & se(\exists c \in C)[classif_{AM}(c) = r \wedge c\ eh\ uma\ VP] \\ 0 & qualquer\ outra\ coisa \end{cases}$$

A precisão de uma dada candidata classificada r é definida na equação 2.7. A função $ehVP(C,r)$ corresponde ao número de VP classificadas de r em C dividido pelo número total de candidatas cuja classificação é menor ou igual a r em C.

Existem outras medidas de avaliação como a Cobertura, que por sua vez coloca no cálculo as candidatas no texto que não foram retornadas e a medida F que é um balanço entre a Precisão e a Cobertura, mas essas não serão usadas neste trabalho.

Capítulo 3

TRABALHOS RELACIONADOS

A proposta apresentada neste documento tem como foco a descoberta automática de EMs no contexto bilíngue. Assim, este capítulo traz um levantamento dos trabalhos da literatura que realizam a descoberta de EMs em contexto bilíngue divididos em dois grupos: (seção 3.1) os que se baseiam em alinhamento de palavras e (seção 3.2) os que utilizam outras fontes de informação.

3.1 Alinhamento de Palavras

Nesta seção abordaremos os trabalhos que usam alinhamento de palavras para realizar a descoberta das EM de forma bilíngue.

3.1.1 Zarrieß e Kuhn (2009)

Zarrieß e Kuhn (2009) trazem um estudo com traduções de expressões verbais (incluindo *light verb constructions*, LVC) no córpus Europarl[1] (inglês e alemão) com base na premissa de que uma grande variação de padrões de EMs podem ser identificados em traduções que mostram uma correspondência entre apenas um item lexical na língua fonte e um grupo de itens lexicais na língua alvo. Para tanto, o método proposto pelos autores descobre EMs usando alinhamento de palavras, gerado pelo o Giza++ (OCH; NEY, 2003), e informações dos *filtros sintáticos*.

O córpus foi analisado sintaticamente pelo *Maltparser* para a etiquetagem das classes gramaticais e dependências. Assim, cada palavra recebe uma indicação de relação sintática com outras palavras, além da indicação de quem é o núcleo, ou seja, quem é a palavra de maior importância naquele sintagma. A análise de dependência é bastante útil para tratar as candidatas

[1] http://www.statmt.org/europarl/

a EMs não contíguas, pois com ela é possível eleger como membros das candidatas apenas as palavras que possuem relações de dependência diretas.

Após a identificação de todos os alinhamentos *um-para-muitos* as dependências sintáticas foram usadas para expandir os alinhamentos, ou seja para inserir algum elemento alvo que deveria ter sindo inserido no alinhamento. Alinhamentos um-para-muitos ocorrem quando uma palavra simples no texto de um lado (fonte ou alvo) foi alinhada com duas ou mais palavras do texto traduzido. Para isso, verificou-se as palavras do lado alvo que tinham dependência com as palavras alinhadas e usou-se a medida *Dice* para calcular a sua associação com o elemento fonte.

Para a avaliação, foram selecionados 50 verbos aleatórios que ocorreram no córpus Europarl, e foi feita a extração das suas traduções no córpus alvo. Esse conjunto de 50 verbos produziu 1.592 alinhamentos *um-para-muitos*, os quais foram reduzidos para 1.302 após a aplicação de um filtro que considerava variações morfológicas. Em linhas gerais, quase 60% das candidatas a EM foram consideradas de fato EMs, o resto foi consideradas paráfrases ou alternâncias. De todas as traduções recuperadas, quase 83% das correspondências entre a fonte e o alvo eram traduções corretas.

Zarrieß e Kuhn (2009), portanto, realizam a descoberta de EMs em córpus paralelos com base em alinhamento de palavras e análise sintática por dependência. Diferentemente desse trabalho, nossa proposta é não utilizar tais recursos externos (alinhador de palavras e analisador sintático por dependência), tornando o processo mais simples e menos dependente/suscetível a recursos externos.

3.1.2 Caseli et al. (2010)

Nesse trabalho, Caseli et al. (2010) desenvolveram um método para descoberta de EMs em um córpus paralelo português-inglês que era uma versão inicial do córpus FAPESP usado nos experimentos apresentados neste documento. No trabalho de Caseli et al. (2010), a versão do córpus FAPESP usada tinha menos sentenças e *tokens* do que a versão atual utilizada neste trabalho. O córpus paralelo contava com 17.397 sentenças e 494.391 *tokens* para o português e 532.121 *tokens* para o inglês. Para tanto, o primeiro passo foi alinhar o córpus em nível de sentença usando a ferramenta TCAalign, desenvolvida com base em (HOFLAND, 1996), e realizar a etiquetação morfossintática por meio dos etiquetadores do sistema de tradução automática Apertium (ARMENTANO-OLLER et al., 2006). Por fim, o córpus foi alinhado lexicalmente usando Giza++ (OCH; NEY, 2003).

A descoberta de EM foi realizada em duas etapas: (1) alinhamento e (2) filtragem. Primeiramente buscou-se alinhamentos assimétricos, ou seja, um alinhamento *um-para-muitos*, como no trabalho de Zarrieß e Kuhn (2009), ou *muitos-para-um*. Por exemplo, um caso real de ocorrência no córpus foi o das palavras *derrames* e *cerebrais* que estavam alinhadas com a palavra única *strokes*.

As listas geradas na primeira etapa foram filtradas usando padrões definidos manualmente com o intuito de eliminar candidatas que não teriam características de EMs. Adicionalmente, foi aplicado um limiar de ocorrências no córpus. Por exemplo, candidatas que possuíam pronomes foram eliminadas da lista. Assim, por exemplo, a lista para o inglês originalmente gerada contava com 37.267 candidatas. Usando o filtro, 27.402 candidatas foram eliminadas por se encaixarem nos padrões. Das candidatas restantes, 8.609 foram eliminadas pois não apareciam pelo o menos 2 vezes no córpus. As 1.256 candidatas restantes passaram por um processo de avaliação que será explicado a seguir.

Na avaliação, foi feita uma busca automática das candidatas em dois dicionários de referência em cada uma das duas línguas. Toda vez que uma candidata era encontrada em algum recurso, a candidata era considerada uma EM, ou seja, uma verdadeira positiva. As candidatas que não foram encontradas nos dicionários passaram por uma anotação manual por humanos. Os anotadores também avaliaram manualmente a qualidade das traduções obtidas automaticamente.Um total de 317 candidatas (25% do total) foram encontradas nos dicionários de referência e suas traduções para o português verificadas por anotadores, uma vez que os dicionários de referência não possuíam tradução. Toda a vez que uma candidata (em inglês) não estava em algum dos dicionários, ela era colocada para avaliação de dois anotadores que não só analisavam se a candidata era uma EM, como também avaliavam a tradução em português. As 939 (75% restantes) candidatas que não foram encontradas nos dicionários, foram avaliadas manualmente por dois anotadores com um índice de discordância de 11%, sendo o valor do coeficiente de concordância κ de 0,768 para anotação de EMs e de 0,761 para traduções.

Foram realizados cálculos de precisão para definir a porcentagem de candidatas apontadas como EMs (positivas verdadeiras) dentro da lista de 1.256. Como as candidatas encontradas nos dicionários já se caracterizam como EMs sem a necessidade de anotadores, adicionou-se ao cálculo 302 candidatas que foram consideradas EMs pelos dois anotadores, totalizando 619 positivas verdadeiras das 1.256 candidatas mostrando uma precisão de 49,28%. Além das 619 EMs já avaliadas, outras 144 candidatas foram consideradas EMs por apenas um dos anotadores e adicionadas ao cálculo anterior. No final 763 foram consideradas EMs totalizando uma precisão de 60,75%. As traduções avaliadas foram apenas das candidatas consideradas verdadeiras

EMs. A avaliação que levou em conta todos os tipos de tradução de todas as candidatas consideradas EMs alcançou uma precisão entre 46,08% a 54,87%. Já para a avaliação das traduções mais frequentes, a precisão foi de 58,61% a 66,92%.

Assim como Zarrieß e Kuhn (2009), Caseli et al. (2010) também realizam a descoberta de EMs em córpus paralelo com base em alinhamento de palavras. Contudo, nesse caso, ao invés de análise sintática por dependência como um passo complementar ao alinhamento ou autores utilizaram a filtragem por padrões morfossintáticos. Nossa proposta tem em comum com esse trabalho o par de idiomas utilizado (português-inglês) e o córpus, uma vez que utiliza uma versão estendida (atual) do córpus FAPESP.

3.1.3 Tsvetkov e Wintner (2012)

Tsvetkov e Wintner (2012) usam o alinhamento de palavras para a descoberta de candidatas e suas respectivas traduções em um córpus bilíngue das línguas inglesa e hebraica. A busca por candidatas não foca em apenas um tipo de EM e o método usado para detecção das traduções consiste em observar quando candidatas do córpus fonte não estão alinhadas *um para um* no córpus alvo. Após aplicar o método e obter um dicionário com as candidatas, as mesmas foram incorporadas em um sistema de tradução automática e a avaliação foi realizada de forma extrínseca.

Os autores usam um córpus paralelo pequeno (TSVETKOV; WINTNER, 2010) composto por 19.626 sentenças, a maioria proveniente de jornais, totalizando 271.87 *tokens* em inglês e 280.508 *tokens* em hebraico. Nesse trabalho, os autores também usaram o MILA córpus (ITAI; WINTNER, 2008), com 46.239.285 *tokens*, que é o córpus monolíngue na língua hebraica. Para o inglês foi usado o *Google's Web IT corpus* (BRANTS; FRANZ, 2006). Os córpus monolíngues, tanto pra o inglês como para o hebraico, foram usados para calcular as estatísticas de distribuição de sequências das palavras de forma separada. Para realizar o alinhamento, os autores usaram o Giza++ (OCH; NEY, 2003). Então, um dicionário bilíngue com 78.313 palavras e suas traduções foi usado para verificar todos os alinhamentos do tipo *uma para um* obtidos automaticamente pelo Giza++.

Diferentemente do trabalho de Zarrieß e Kuhn (2009), a abordagem desse trabalho não se limita apenas às buscas de alinhamento do tipo *um para muitos*, mas também analisa os alinhamentos *um para um* e propõe uma análise de estudo com foco exatamente nos desalinhamentos do córpus. Em linhas gerais, uma vez que os alinhamentos *um para um* encontrados em um dicionário bilíngue eram eliminados do texto, o que restava eram exatamente as candidatas que estavam desalinhadas. Para essas candidatas, usou-se o PMI^K (uma variação do PMI descrito

na seção 2.2) como medida de associação para as candidatas. As candidatas foram analisadas de forma separada fazendo uma busca delas nos córpus monolíngues para o hebraico e para o inglês. Candidatas cuja pontuação de PMI^K era inferior a um determinado limiar, e candidatas que não ocorreram nenhuma vez no córpus monolíngue, eram eliminadas. Como resultado, as candidatas então descobertas incluíam principalmente nomes próprios, locais geográficos, compostos nominais e combinações de substantivos e adjetivos

A avaliação foi baseada na precisão em k, como descrito na seção 2.5.3. As 100 primeiras candidatas ranqueadas de acordo com o PMI^K foram analisadas por dois anotadores e constatou-se que 99 delas eram, de fato, EMs (ou seja, uma precisão de 99% considerando-se as top-100). Para complementar essa avaliação inicial, outras três avaliações mais minuciosas foram realizadas, como explicado a seguir.

As traduções encontradas no córpus paralelo que formavam sequências contíguas e não eram formadas por mais de 4 palavras, foram então incorporadas em dicionário bilíngue para serem inseridas em um tradutor automático. Essas traduções foram obtidas por uma abordagem simples de relacionar EMs descobertas com suas traduções no córpus alvo. As traduções também foram geradas a partir das EMs descobertas anteriormente e não apenas as 100 avaliadas. Os pares de tradução encontrados totalizaram 3.310 EMs e foram incorporados então com as 78.313 entradas que o tradutor automático Hebraico-Inglês já possuía em seu dicionário bilíngue. Foram usadas então as medidas BLEU (PAPINENI et al., 2002) e Meteor (DENKOWSKI; LAVIE, 2011) para avaliar as traduções do tradutor automático com seus dicionários originais e aumentados. Respectivamente, a medida BLEU aumentou de 13,69% para 13,79% e o Meteor de 33,38% para 33,99%, indicando que as EM podem ajudar na tradução automática.

Com uma visão oposta à dos trabalhos anteriores (ZARRIESS; KUHN, 2009; CASELI et al., 2010), que partiam de alinhamentos *um para muitos* ou *muitos para um* como indícios de candidatas a EM, Tsvetkov e Wintner (2012) encontram primeiro os alinhamentos diferentes de *um para um* excluindo-os, e então olham para os restantes como possíveis candidatas a EM. Infelizmente, a comparação direta das estratégias empregadas por esses autores não é possível porque as metodologias de avaliação são muito distintas. Contudo, vale ressaltar que embora o resultado da avaliação intrínseca apresentado por esse trabalho (99% de precisão nas top-100) impressione, o resultado da avaliação extrínseca é menos impressionante (apenas 0,1 e 0,6 pontos de melhora em BLEU e Meteor, respectivamente, na avaliação da saída da tradução automática quando as listas de EMs foram adicionadas aos recursos de um tradutor automático).

3.1.4 Smith (2014)

O trabalho de Smith (2014) tem como objetivo descobrir construções verbo-partícula a partir de um córpus paralelo inglês-espanhol. Nesse trabalho, o córpus foi alinhado por palavras gerando, assim, as correspondências entre as línguas. A partir desses alinhamentos, foram mantidos apenas aqueles casos em que a candidata fonte é um verbo seguido de um complemento e a sua correspondente no espanhol é pelo menos um verbo. O trabalho se propõe a extrair construções verbo-partículas tendo como base (CASELI et al., 2010), que conseguiu bons resultados para *phrasal verbs* (verbos frasais).

A metodologia desse estudo foi adaptada de (CASELI et al., 2010) e consiste em 4 estágios: (1) pré-processamento do córpus com etiquetação morfossintática feita pelo TreeTagger (SCHMID, 1994) e alinhamento por palavras usando o Giza++ (OCH; NEY, 2003) , (2) descoberta de candidatas a EM, (3) filtragem e (4) agrupamento dessas candidatas. O córpus paralelo utilizado era composto por uma coleção de legendas de séries e filmes do site *Open Subtitles*[2] (TIEDEMANN, 2012) nas línguas espanhola e inglesa, com 342.833.112 *tokens* em inglês e 299.880.802 em *espanhol*.

As candidatas foram extraídas considerando-se todas as ocorrências onde os verbos no lado fonte (inglês) fossem do tipo verbo-partícula, ou seja, eram formados por uma palavra marcada como verbo seguido de uma palavra marcada como partícula e estavam alinhados no lado alvo (espanhol) por uma palavra marcada como verbo. Nesse caso, por exemplo, *come on* (em inglês) seria extraído porque estaria alinhado com *vamos* (em espanhol); da mesma forma que *derrames cerebrais* era extraído por estar alinhado com *strokes* em (CASELI et al., 2010). À medida que essas ocorrências apareciam no mínimo 5 vezes, elas eram extraídas e tomadas como candidatas. Após essa extração, a lista de candidatas gerada contendo 18.186 entradas foi agrupada por seus lemas, resultando em uma lista menor contendo 6.833 entradas, com 1.424 candidatas a construções verbo-partícula distintas.

A avaliação intrínseca foi realizada considerando-se 100 candidatas selecionadas de forma randômica dentre as 1.424. Das 100 candidatas, um falante nativo de inglês validou 94 como corretas (ou seja, 94% de precisão nessa amostra). Das 94 válidas, pesquisou-se em dicionários online para saber se elas apareciam por lá, resultando em que 80 estavam presentes (ou seja, em uma avaliação realizada apenas com base em um *gold standard* constataria-se uma precisão de 80%). Das 94 válidas, 75 delas possuíam uma tradução correta para o espanhol, julgada por um falante nativo de espanhol (ou seja, a precisão na definição das traduções corretas foi de 79,8%).

[2]https://www.opensubtitles.org/

Os trabalhos abordados até agora usam o alinhamento de palavras como ferramenta principal para descobrir EMs em contexto bilíngue. Como vimos, nem sempre verbos, colocações ou expressões idiomáticas de um lado fonte estarão alinhados com o mesmo número de palavras no lado alvo. Essa diferença de número de palavras no alinhamento pode ser encarada como um problema, uma vez que tende-se a esperar que traduções de um lado sejam traduzidas como expressões do outro, ou como uma estratégia para a descoberta de EMs, como vimos nos trabalhos citados nesta seção. Zarrieß e Kuhn (2009) criam um método para olhar não apenas para as ocorrências de números diferentes de palavras alinhadas, mas para verificar se algumas palavras na verdade não pertenciam ao alinhamento. Caseli et al. (2010) e Smith (2014) seguem estratégia similar baseada na diferença no número de palavras alinhadas e fazem um estudo mais aprofundado exatamente nesse *desalinhamento*. Da mesma forma, Tsvetkov e Wintner (2012) olham para o texto e removem as ocorrências em que as palavras são alinhadas *um para um* analisando o restante mais a fundo na busca por EMs. Para a proposta aqui apresentada, optamos por não usar o alinhamento de palavras, uma vez que novas estratégias têm se mostrado mais promissoras, como relatam os trabalhos apresentados na subseção a seguir e os experimentos iniciais apresentados no Capítulo 5.

3.2 Outras Abordagens

A seguir são apresentados alguns métodos de descoberta automática de EMs, no contexto bilíngue que não utilizam a informação de alinhamento de palavras como base para suas estratégias.

3.2.1 Bouamor, Semmar e Zweigenbaum (2012)

Nesse trabalho EMs são extraídas de um córpus paralelo francês-inglês e os resultados são incorporados em um tradutor automático para tentar melhorar o desempenho das traduções. A descoberta é feita com base em padrões morfossintáticos, considerando apenas as candidatas frequentes no córpus (filtro por frequência mínima de ocorrência), e a definição da melhor tradução é feita com base em um modelo de espaço vetorial (SALTON; WONG; YANG, 1975).

Para a extração de pares de EMs fonte e alvo, foram considerados n-gramas, onde o n correspondia a 2, 3 e 4, que coincidiam com padrões morfossintáticos definidos manualmente. Os padrões eram combinações de adjetivo, substantivos e preposições tais como adjetivo-substantivo, substantivo-substantivo, substantivo-preposição-substantivo. O córpus paralelo foi processado com a plataforma CEA LIST Multilingual Analysis (LIMA) (BESANÇON et al., 2010) para

lematizar e etiquetar morfossintaticamente os textos no lado fonte e alvo. As dimensões e características do córpus paralelo não foram informados.

Para realizar as comparações das candidatas alvo com as candidatas fonte, o trabalho usa um modelo de espaço vetorial no qual um vetor de tamanho N é associado a cada candidata. O valor N é o número de sentenças do córpus, e cada posição desse vetor é marcada se a candidata aparece na sentença que essa posição representa. Por exemplo, se uma candidata fonte aparece nas sentenças 3 e 4 do córpus fonte, as posições 3 e 4 do vetor são preenchidas com o valor 1 e as posições restantes são marcadas com 0. Tais vetores são gerados para as candidatas fonte e alvo. Em seguida, os vetores são comparados para o estabelecimento das correspondências entre eles. Para tanto, as buscas por correspondências é realizada considerando-se as candidatas fonte mais frequentes primeiro. Essa comparação é feita usando o índice *Jaccard* que calcula a proporção de vezes em que as candidatas fonte e alvo aparecem na mesma sentença. A candidata alvo com melhor índice de Jaccard para uma dada candidata fonte é selecionada como sua tradução e o par de candidatas a EM vai para a lista final. Apesar de não usar alinhadores de palavras como GIZA++, a comparação de vetores implicitamente realiza um alinhamento superficial pois usa a co-ocorrência em sentences paralelas para encontrar as traduções. No entanto, ao invés de alinhar as palavras como no caso do GIZA++, as traduções são obtidas diretamente para os n-gramas candidatos a EM. O trabalho não menciona se a lista final passou por uma avaliação manual.

O método de descoberta de EMs e suas traduções foi avaliado extrinsecamente por meio da integração da lista gerada em um tradutor automático estatístico. A avaliação foi realizada considerando-se os valores de BLEU (PAPINENI et al., 2002) e Meteor (DENKOWSKI; LAVIE, 2011) para o tradutor treinado de 4 formas diferentes: (1) tradutor *baseline* treinado sem qualquer indicação de EM, (2) tradutor treinado usando a lista de EMs como um córpus paralelo gerando um modelo com mais informações, (3) tradutor treinado colocando as EMs na tabela de tradução (para auxiliar diretamente nas traduções) e, por último, (4) tradutor treinado com com um atributo novo na tabela de tradução. Esse atributo é binário e indica se aquela sequência de palavras é uma EM ou não. Essas estratégias são similares às utilizadas por outros trabalhos que tentam incluir EMs em tradutores automáticos (REN et al., 2009; CARPUAT; DIAB, 2010).

Os resultados da avaliação demonstram que os tradutores treinados considerando-se as EMs nas versões (2) e (3) foram melhores do que o *baseline*: 25,94 de BLEU e 29,26 de Meteor para a versão (2) e 25,67 e 29,26 para a versão (3) contra 25,64 e 29,11 do *baseline*. Os resultados mostram que a abordagem (2) teve um aumento de 0,3 e 0,15 para as medidas BLEU e Meteor respectivamente, em relação com o *baseline*. Por outro lado, a versão (2) teve aumentos de

apenas 0,03 e 0,15. Já a estratégia (4) não se mostrou melhor do que o *baseline*, com 22,91 de BLEU contra 27,18 obtido pelo *baseline*, mostrando um resultado inferior.

Algumas das ideias desse trabalho também são exploradas na proposta apresentada neste documento, como o uso de padrões morfossintáticos para a extração de candidatas e o de um modelo de espaço vetorial, juntamente com a co-ocorrência nas sentenças paralelas, para se determinar a correspondência entre candidatas fonte e alvo. A avaliação extrínseca, no entanto, não será explorada no presente trabalho.

3.2.2 Pereira et al. (2014)

O trabalho de Pereira et al. (2014) tem o objetivo de descobrir colocações e expressões idiomáticas a partir de um córpus paralelo japonês-inglês. Diferentemente de trabalhos como os citados na seção 3.1, este trabalho não usa alinhamento de palavras para gerar traduções, mas sim, um dicionário bilíngue. O método proposto usa as traduções do dicionário para identificar as chances de a candidata em japonês ser, de fato, uma colocação/expressão idiomática. O trabalho se baseia na seguinte ideia: se uma dada uma candidata em japonês é traduzida para o inglês e essa tradução não parece ser amplamente usada em inglês, significa que essa expressão em japonês possui fortes evidências de ser uma colocação ou expressão idiomática. A justificativa pelo uso de dicionários ao invés do alinhamento de palavras está fundamentada em dois pontos que tornariam o alinhamento lexical mais suscetível a erros no tratamento desse tipo de EM: (1) o fato de que expressões idiomáticas, em sua maioria, não são traduzidas de forma literal para outra língua e (2) o número de palavras de uma EM fonte não é igual ao número de palavras da sua tradução.

O córpus utilizado nos experimentos foi o *Hiragana Times*[3], um córpus bilíngue japonês-inglês contendo artigos de revistas publicados de 2002 a 2012. O córpus contém um total de 117.492 pares de sentenças paralelas, com 3.949.616 *tokens* para o japonês e 2.107.613 *tokens* para o inglês. As candidatas foram extraídas pelo analisador sintático por dependência Cabocha (KUDO; MATSUMOTO, 2002). As candidatas extraídas foram construções substantivo-verbo. Após a extração, as candidatas foram ordenadas por frequência de ocorrência.

O dicionário bilíngue japonês-inglês Edict (BREEN, 1995) contendo 110.424 entradas foi usado para realizar a tradução de cada candidata fonte extraída como descrito anteriormente. O dicionário gerou todas as possíveis combinações de traduções das candidatas em japonês, uma vez que uma palavra na língua alvo pode ter vários significados diferentes quando traduzida

[3] http://www.hiraganatimes.com

literalmente. Traduções que continham mais de 3 palavras foram eliminadas e o dicionário gerou uma média de 4,5 traduções para cada candidata.

Uma vez que as candidatas foram extraídas e as suas traduções geradas, o próximo passo foi gerar medidas de associação para ambos os lados (fonte e alvo). O mwetoolkit (RAMISCH, 2012) foi usado para calcular as medidas de associação para todas as candidatas (japonês e inglês). Os autores escolheram a medida *Dice* para ser usada no método por ter mostrado um melhor desempenho. Para as candidatas em japonês, as medidas de associação foram calculadas usando o córpus bilíngue Hiragana times, mas para as candidatas em inglês usou-se um córpus monolíngue formado por 75.377 artigos em inglês da *Wikipedia* (contendo 9.500.000 sentenças e 247.355.886 tokens) etiquetado morfossintaticamente para facilitar o processo de extrair candidatas usando o mwetoolkit. O próximo passo então foi pegar os valores de *Dice* de cada candidata em japonês e dividir pela média dos valores de *Dice* de suas traduções. O trabalho investigava o pressuposto de que se uma candidata fonte tivesse uma pontuação alta de *Dice* e sua tradução através de um dicionário tivesse uma pontuação baixa de *Dice*, seria indício de que a candidata fonte era uma EM. Após o cálculo, as candidatas foram classificadas pelo resultado da razão: a candidatas com razões altas tinham grandes chances de serem colocações ou expressões idiomáticas e as candidatas com razões baixas tinham chances de serem combinações aleatórias.

Para a geração de um córpus de teste, as 100 candidatas mais frequentes extraídas do córpus bilíngue foram anotadas por falantes nativos de japonês. A tarefa desses anotadores era decidir se cada candidata era uma colocação/expressão idiomática ou uma combinação aleatória de palavras. Para as 100 candidatas anotadas, a concordância entre os anotadores calculada usando o coeficiente estatístico *Fleiss' Kappa* (FLEISS, 1971) foi de 0,4354. Para formar o córpus de teste final, apenas as candidatas em que mais de 3 anotadores concordaram quanto a sua classificação foram adicionadas ao conjunto de testes. O resultado foi um córpus de teste composto por 87 instâncias sendo 36 colocações/expressões idiomáticas e 51 combinações aleatórias. A concordância entre anotadores nesse córpus de teste foi de 0,5427.

A abordagem proposta por Pereira et al. (2014) foi comparada com dois *baselines*: (1) descoberta de EMs usando apenas os valores do *Dice* e (2) descoberta de EMs usando um tradutor estatístico baseado em frases. O tradutor estatístico foi desenvolvido usando o *toolkit* Moses (KOEHN et al., 2007) e obteve uma precisão média nas traduções das candidatas presentes no conjunto de teste de 67%, enquanto que as candidatas classificadas apenas com base no valor de *Dice* obtiveram 54%. O método proposto para esse trabalho obteve uma precisão de 71% o que supera desempenho dos dois *baselines*. Isso significa que o método proposto de cálculo da

razão entre as pontuações de Dice das candidatas e suas traduções melhora o método simples baseado em Dice, pois muitas candidatas que possuem poucas chances de serem colocações, por exemplo, receberam valores de Dice muito grandes.

O método de Pereira et al. (2014) traz um ponto também explorado na proposta apresentada neste documento: comparar os valores de medidas gerados para candidatas fonte e alvo na busca pelos melhores pares de EMs e traduções. No caso do trabalho descrito nesta subseção, o valor comparado foi o de *Dice*, no caso da proposta deste trabalho investiga-se o uso de similaridade de cosseno e outras medidas de associação como detalhado no Capítulo 4.

3.2.3 Garcia, García-Salido e Alonso-Ramos (2017)

O trabalho de Garcia, García-Salido e Alonso-Ramos (2017) realizam a descoberta de EMs do tipo colocações em córpus paralelo gerando uma lista de candidatas para as línguas inglesa, portuguesa e espanhola e usando os mesmos córpus para o treino de vetores bilíngues de palavras. Em linhas gerais, primeiramente, as candidatas são extraídas de forma monolíngue separadamente para as três línguas e, então, os vetores de palavras são usados para relacionar as candidatas de acordo com sua similaridade.

As candidatas foram extraídas de um subconjunto do córpus paralelo OpenSubtitles2016 (LISON; TIEDEMANN, 2016). Por ser um córpus de legendas de filmes e séries de televisão, foram selecionadas 13.017.016 sentenças que apareciam nas três línguas, totalizando cerca de 91 milhões de *tokens* para o espanhol e português e cerca de 98 milhões de *tokens* para o inglês. Antes da extração monolíngue, os córpus foram lematizados e etiquetados morfossintaticamente usando o método de Garcia e Gamallo (2015). Além disso os córpus nas três línguas foram processados pelo MaltParser (NIVRE; HALL; NILSSON, 2006) para receber informações sintáticas. O resultado disso é um córpus analisado com etiquetas morfossintáticas, lemas e dependências sintáticas.

Os vetores de palavras foram treinados com o Multivec (BÉRARD et al., 2016) usando os córpus paralelos na forma lematizada. Os vetores bilíngues foram gerados para os seguintes pares de idiomas: espanhol-inglês, espanhol-português e português-inglês.

Como já mencionado, esse trabalho tem o foco na descoberta de colocações. Para tanto, foram definidos três tipos de colocações: adjetivo-substantivo, substantivo-substantivo e verbo-objeto. As candidatas foram extraídas com base em padrões morofossintáticos. As candidatas extraídas foram classificadas usando as medidas *Mutual Information*, por funcionar muito bem em córpus grandes (PECINA, 2010), e *t-score*, que se mostrou útil para extração de colocações

pois leva em conta a alta frequência dos bigramas (KRENN; EVERT et al., 2001). Foram extraídas apenas as candidatas com uma frequência no córpus de no mínimo 10 ocorrências, e os limiares para as medidas de associação foram: *Mutual Information* maior ou igual a 3 e *t-score* maior ou igual 2.

Após a etapa de extração monolíngue aplicada aos lados fonte e alvo dos textos paralelos, a próxima etapa foi relacionar as candidatas do lado fonte com as candidatas do lado alvo. Nesse processo, para relacionar as candidatas de um idioma fonte F com as candidatas de um idioma alvo A utilizou-se: as listas de candidatas no idioma F, as listas de candidatas no idioma A e os vetores bilíngues F-A. Por exemplo, para relacionar as candidatas espanhol-inglês, considerou-se as listas de candidatas em espanhol, as listas de candidatas em inglês e os vetores bilíngues espanhol-inglês. Os autores utilizam os termos *base* e *colocador* para se referir aos dois elementos da colocação (MEL'ČUK, 1998). Assim, para cada candidata fonte, sua base é selecionada e os 5 lemas mais similares na língua alvo são recuperados com base nos vetores bilíngues. Por exemplo, para o espanhol *tremendo lío* (tremenda bagunça), a base da candidata é a palavra *lío*. Com o modelo bilíngue, são encontradas palavras alvo similares como *trouble* e *mess* (que significam bagunça ou problema). Em seguida, as candidatas alvo que contêm um desses lemas são recuperadas, por exemplo: *little trouble*, *deep trouble*, *huge mess*, *fine mess*. Em seguida, os *colocadores* das candidatas são comparados para medir sua similaridade (chances de uma palavra ser a tradução da outra), no exemplo: *tremendo* com *little*, *deep*, *huge* e *fine*. Se o valor de similaridade for maior que 0,65 (valor definido empiricamente no trabalho) e se o *colocador* alvo estiver entre os 5 lemas mais similares do *colocador* fonte, então as duas candidatas são extraídas e associadas a um *score* que é o valor médio da similaridade entre as *bases* e os *colocadores* nas duas línguas.

Esse método foi comparado com um método que utilizava dicionários bilíngues para estabelecer a relação das candidatas fonte e alvo (*baseline*). Foram utilizados os dicionários bilíngues do *Apertium* (FORCADA et al., 2011) com: 14.364 entradas espanhol-português e 34.994 entradas espanhol-inglês. Para o português-inglês, utilizou-se um dicionário obtido automaticamente a partir de outros dois léxicos (não pelo *Apertium*) contendo 9.160 entradas.

Para a comparação dos dois métodos, 100 pares alinhados de candidatas foram selecionados aleatoriamente para cada método, par de línguas e padrão de extração, totalizando 9 listas. Dois anotadores avaliaram as entradas dessas listas considerando como corretas apenas aquelas para as quais os lados fonte e alvo eram colocações e a tradução estava correta. Em termos de concordância entre anotadores (o artigo não informa qual foi a medida usada), o *baseline* alcançou 92% e o método proposto pelo trabalho 93%. As candidatas em que os anotadores

não concordaram foram excluídas das avaliações.

Na avaliação final, o método *baseline* obteve 91,7% de precisão no pior cenário (pois cada padrão e línguas tinha sua própria lista, como mencionado anteriormente). Entretanto os valores de revocação ficaram entre 1,2% e 10,5%, o que acarretou uma medida F em torno de 4,7%. Já para o método proposto no trabalho, os valores de precisão ficaram um pouco mais baixos, entre 83,9% e 92,9%. Apesar dos valores de precisão terem sido inferiores aos do *baseline*, os valores de revocação foram superiores, chegando a valores como 66% para alguns casos. Como resultado, a medida F do método proposto apresenta valores muito maiores: entre 40% a 77%. Analisando os números, é possível perceber que mesmo que os dicionários sejam bons em termos de precisão, uma vez que têm um alto grau de qualidade, os modelos baseados na distância entre os vetores de palavras mostram-se bastante efetivos no cálculo da similaridade entre palavras aumentando, assim, a cobertura do método.

Em relação à proposta de técnicas de descoberta apresentadas neste documento, Garcia, García-Salido e Alonso-Ramos (2017) é, sem sombra de dúvida, o trabalho mais similar. Para o nosso trabalho, também propomos usar o método de recuperação de palavras similares com base nos vetores bilíngues de palavras aplicados para medir o grau de similaridade entre duas candidatas. Esse método será explicado no Capítulo 4.

Podemos verificar que nesta seção as abordagens para descoberta bilíngue de EMs se diversificam ao longo do tempo. O método de Bai et al. (2009) gera traduções através do texto alvo usando medidas de associação. Já Bouamor, Semmar e Zweigenbaum (2012) usam uma abordagem de extrair uma candidata no lado fonte e procurar pela sentença paralela. Para Pereira et al. (2014) a ideia foi gerar possíveis traduções através de um dicionário bilíngue e aplicar uma medida de associação para essas possíveis traduções. Por último Garcia, García-Salido e Alonso-Ramos (2017) geram traduções através dos vetores de palavras. Como optamos até o momento por não usar alinhamento de palavras, é possível dizer que esses métodos são os mais parecidos com o nosso trabalho. A nossa abordagem faz uma busca por candidatas relacionando sentenças como Bouamor, Semmar e Zweigenbaum (2012) e faz uma busca por traduções através dos vetores de palavras como Garcia, García-Salido e Alonso-Ramos (2017).

As principais características dos trabalhos relacionados citados neste capítulo podem ser visualizadas no quadro comparativo da Tabela 3.1.

Tabela 3.1: Trabalhos Relacionados

Autor	Tipo EM	Método	Alinhamento	Línguas	Córpus
Zarrieß e Kuhn (2009)	expressões verbais	Identificação de alinhamento *um para muitos* e traduções	Sim	Inglês e Alemão	Europarl
Caseli et al. (2010)	Todos	Alinhamento *um para muitos* e suas traduções	Sim	Português e Inglês	Córpus Bilíngue Fapesp
Smith (2014)	Contrução Verbo-Partícula	Alinhamento de sentenças e padrão com verbos	Sim	Inglês e Espanhol	Open Subtitles
Tsvetkov e Wintner (2012)	Todos	Alinhamento diferente de *um para um* e suas traduções	Sim	Inglês e Hebraico	MILA corpus Google's Web IT corpus
Bouamor, Semmar e Zweigenbaum (2012)	Compostos nominais	Padrões morofossintáticos Vetores *Jaccard*	Não	Francês - Inglês	Textos da Wikipédia
Pereira et al. (2014)	Colocações Expressões idiomáticas	Analisador Sintático por dependência Medida de associação	Não	Japonês e Inglês	Hiragana Times
Garcia, García-Salido e Alonso-Ramos (2017)	Colocações	Medidas de associação Vetores bilíngues de palavras	Não	Várias	Open Subtitles

Capítulo 4

BIDIMWETOOL

Este trabalho teve como objetivo a descoberta de EMs em textos paralelos sem utilizar alinhamento de palavras. Batizamos esta nossa proposta de *Bilingual Discovery MWE Tool* ou, apenas, *BiDiMWETool*. Desenvolvemos dois métodos para a descoberta bilíngue de EMs. O primeiro deles extrai candidatas a EM nas duas línguas e, depois, tenta descobrir as relações de tradução entre elas. O segundo, faz a extração de candidatas apenas no idioma fonte, e depois tenta descobrir suas traduções nas sentenças paralelas.

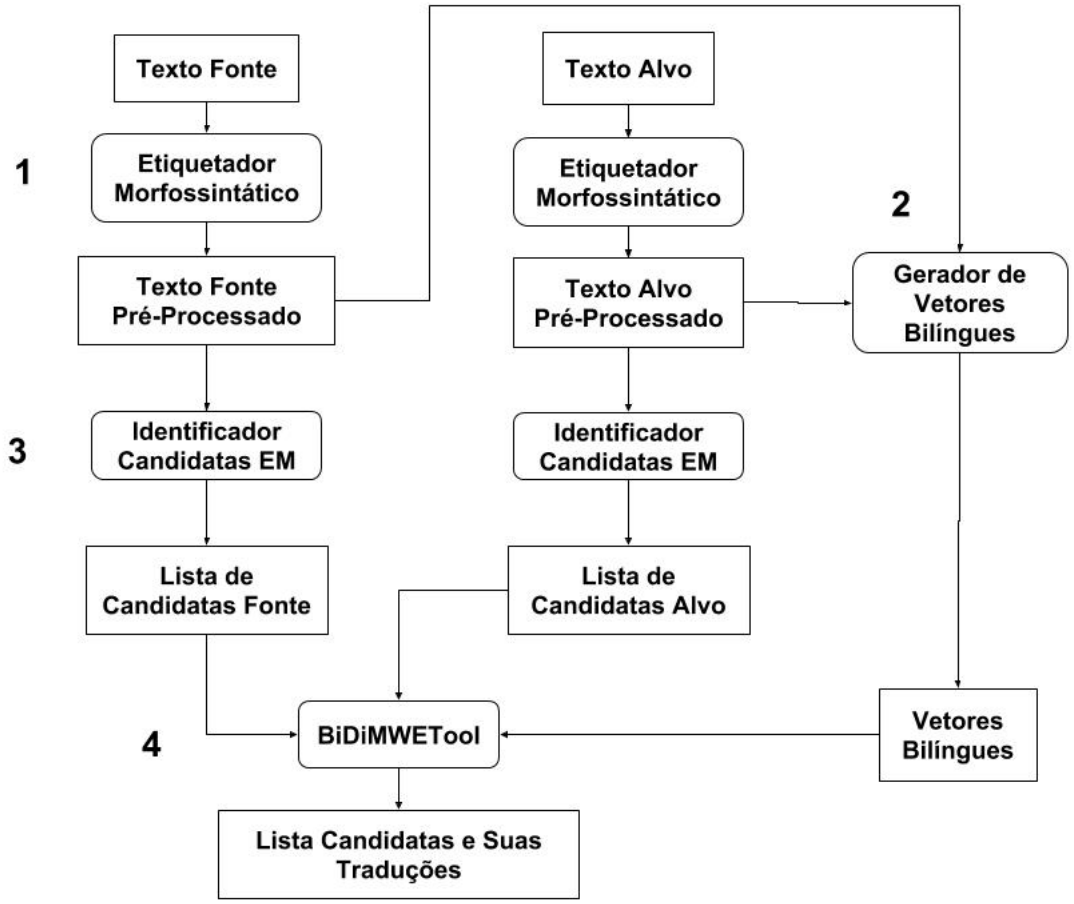
A arquitetura do *BiDiMWETool* é baseada na descoberta independente de candidatas em cada lado do córpus paralelo com o uso de padrões morfossintáticos. É importante ressaltar que cada vez que mencionamos o termo *descoberta monolíngue*, estamos falando da descoberta de EMs usando as medidas de associação descritas na seção 2.2. Em seguida, realiza-se a descoberta bilíngue através do pareamento e filtragem de uma ou das duas listas de candidatas usando vetores de palavras bilíngues, medidas de associação e co-ocorrência em sentenças paralelas. Em linhas gerais, o resultado da ferramenta são duas ou três listas de candidatas: uma com as candidatas a EM para português, outra com as candidatas a EM para inglês (apenas no primeiro método) e a última com o produto da descoberta bilíngue com candidatas fonte associadas a candidatas (ou palavras) alvo, indicando as chances das candidatas (ou palavras) alvo serem a tradução das candidatas fonte.

As duas próximas seções apresentam a arquitetura geral dos dois métodos de descoberta bilíngue desenvolvidos neste trabalho. As técnicas e ferramentas usadas serão apresentados no próximo capítulo.

4.1 Primeiro método de descoberta bilíngue

A Figura 4.1 ilustra a arquitetura proposta para o primeiro método de descoberta bilíngue desenvolvido neste trabalho. Essa arquitetura representa a sequência de tarefas propostas para o primeiro método de descoberta. Os retângulos com cantos arredondados representam processos e os demais retângulos representam recursos.

Figura 4.1: Arquitetura do primeiro método de descoberta bilíngue automática de EMs a partir de córpus paralelo no *BiDiMWETool*.



As quatro etapas do primeiro método proposto são descritas a seguir:

1. O córpus paralelo é fornecido como entrada. Esse córpus paralelo contém pares de texto fonte e texto alvo, alinhados por sentenças. Um exemplo de sentenças paralelas do córpus fornecido como entrada é apresentado na Tabela 4.1. Trata-se de sentenças presentes no córpus FAPESP (AZIZ; SPECIA, 2011) usado em nossos experimentos. Para possibilitar a busca por candidatas a EM que satisfaçam determinados padrões morfossintáticos, precisamos que o córpus passe por um processo prévio de etiquetação morfossintática realizado por uma ferramenta computacional.

O resultado desse processo é o córpus anotado de forma automática com suas respectivas classes gramaticais e suas formais canônicas. Por exemplo, a palavra *realizei*, após a etiquetação morfossintática, seria anotada com o lema *realizar*, a categoria gramatical *verbo* e demais traços morfológicos indicativos de tempo e pessoa, por exemplo. Essas informações facilitam a formulação de padrões. Por exemplo, se quisermos pesquisar todas as vezes que o verbo *realizar* apareceu seguido de um substantivo, a busca pode ser especificada com base no lema e retornar casos em que apareceram ocorrências do verbo flexionado: *realizei*, *realizaram*, etc.

Tabela 4.1: Exemplo de córpus paralelo de entrada para o *BiDiMWETool*

Texto fonte (português)	**Texto alvo (inglês)**
E o que poderia articulá-los.	What could link them?
- Uma nova linguagem que permitisse às diversas ciências se comunicar com rapidez e clareza.	- A new language to allow the various sciences to communicate with each other fast and clearly.

Nessa primeira etapa, cada um dos textos que compõem o córpus paralelo passa, separadamente, pelo processo de etiquetação morfossintática e lematização. Como essa é uma etapa de pré-processamento, o usuário é livre para escolher qualquer ferramenta disponível para etiquetação e lematização. Um exemplo de sentenças paralelas processadas por um etiquetador morfossintático é apresentado na Tabela 4.2.

Tabela 4.2: Exemplo de sentenças paralelas após o processo de etiquetação morfossintática e lematização

Texto fonte (português)	**Texto alvo (inglês)**
Uma DET um | nova ADV novo | linguagem SUBS linguagem | que PRON que | permitisse VERBO permitir | às PREP a+as | diversas ADV diverso | ciências SUBS ciência | se PRON se | comunicar VERB comunicar | com PREP com | rapidez SUBS rapidez | e CONJ e | clareza SUBS clareza | . PONT .	A DET a | new ADJ new | language SUBS language | to PREP to | allow VERB allow | the DET the | various ADV various | sciences SUBS science | to PREP to | communicate VERB communicate | with PREP with | each ADV each | other ADV other | fast SUBS fast | and CONJ and | clearly ADV clearly | . PONT .

O exemplo apresentado nessa tabela contém um par de sentenças paralelas do córpus FAPESP lematizadas e etiquetadas morfossintaticamente. No exemplo, as palavras são separadas umas das por barras verticais (|) e são formadas pelas triplas: forma, etiqueta morfossintática, lema. A palavra *linguagem*, por exemplo, foi classificada como substantivo e sua forma lematizada tem a mesma forma. Já a palavra *permitisse*, foi classificada como verbo e sua forma lematizada é *permitir*. É importante ressaltar que etiquetadores

possuem seus próprios *padrões de etiquetas*. No nosso exemplo, adaptamos as mesmas para fins de melhor entendimento das classes gramaticais.

2. Os pares de textos paralelos são, também, usados pela ferramenta responsável por gerar vetores de palavras bilíngues. Os vetores bilíngues são gerados a partir dos pares de textos paralelos apenas lematizados e sem etiquetação morfossintática. O resultado desse processo é um arquivo que é usado para consulta pelo BiDiMWEToolkit. Nesse arquivo existe o mapeamento e a relação das palavras em português com as palavras em inglês em um espaço vetorial, como vimos na seção 2.4. Exemplo dos resultados de consultas feitas a esses vetores bilíngues podem ser vistos na Tabela 4.3.

Tabela 4.3: Exemplo de resultados para consultas feitas aos vetores bilíngues

Palavra fonte (português)	**Palavras alvo (inglês) mais próximas**
verificar	verify, ascertain, check, detect, prove
trabalho	work, investigation, review, field, conclusion
assunto	subject, theme, issue, topic, question

3. Os textos etiquetados morfossintaticamente em cada língua são fornecidos, separadamente, como entrada para o identificador de candidatas a EM, que gera uma lista de EMs candidatas. Essa ferramenta faz a descoberta monolíngue no córpus fonte e alvo, separadamente, baseada nas informações de lema e etiquetação morfossintática usando as medidas de associação já apresentadas na seção 2.2 deste trabalho. A partir dessas listas de EMs fonte e alvo candidatas, o usuário tem a possibilidade de escolher quais candidatas estão melhor classificadas de acordo com alguma medida calculada pelo *toolkit*.

4. A partir das listas de candidatas fonte e alvo e dos vetores bilíngues, o *BiDiMWETool* gera a lista de candidatas a EM e suas traduções. Para tanto, ele faz uma associação entre as duas listas monolíngues (lista fonte e alvo) com base nas seguintes informações: número de vezes em que as candidatas fonte aparecem na mesma linha que certas candidatas alvo (Probabilidade) e com base nos vetores bilíngues, o grau de semelhança que duas candidatas possuem em um espaço vetorial (Cálculo da Similaridade). Explicamos esses passos a seguir:

   - Probabilidade condicional (P): uma vez que temos as candidatas pareadas (fonte e alvo), realizamos o cálculo da probabilidade do par, $P(a_j|f_i)$, definida pela equação

   $$P(a_j|f_i) = \frac{c(f_i, a_j)}{c(f_i)}$$

   a probabilidade $P(a_j|f_i)$ é dada pelo número de vezes que uma candidata alvo (a_j)

apareceu na mesma linha que a candidata fonte (f_i), $c(f_i, a_j)$, em relação ao número de ocorrências da candidata fonte, $c(f_i)$. Se, por exemplo, o pareamento a candidata *dar ímpeto* apareceu 5 vezes durante a extração monolíngue e foi pareada 3 vezes com a candidata alvo *give impetus*, o cálculo da probabilidade será 3/5,

- Similaridade (*Sim*): nós também calculamos a *similaridade distributiva multilíngue* entre os pares de candidatas fonte (f_i) e alvo (a_j). Esse cálculo faz uso de vetores bilíngues de palavras pré-treinados, assumindo que as palavras que são as traduções uma da outra próximas nesse mesmo espaço semântico. Como cada par de candidatas f_i e a_j é composto por m e n palavras, respectivamente, nós usamos a similaridade do cosseno média entre todas as possibilidades $m \times n$ de pares fonte-alvo presentes no espaço semântico:

$$Sim(f_i, a_j) = \frac{1}{m \times n} \sum_{\substack{k=1..m \\ l=1..n}} cos(w_k^{f_i}, w_l^{a_j})$$

onde $w_k^{f_i}$ e $w_l^{a_j}$ são os vetores de palavras para a k-ésima palavra na candidata fonte e a l-ésima palavra na candidata alvo, respectivamente. *cos* é o valor calculado para a similaridade de cosseno de dois vetores. Se em algum momento o resultado da similaridade entre um par (f_i, a_j) for 0, assumimos como valor final o valor de maior similaridade entre as outras candidatas (f_i, a_j), descartando o valor 0 do nosso cálculo.

Por fim, a pontuação do par fonte-alvo é calculada como segue:

- Pontuação final (F): a pontuação final de um par de candidatas fonte-alvo é calculada com base em três medidas: P, t-score (veja seção 2.2) da candidata alvo e *Sim*. Todos os valores calculados para essas medidas são normalizados para que estejam entre 0 e 1. O valor final F é uma combinação log-linear dessas medidas:

$$F(t_j|s_i) = \sum_{f \in \{P, tscore, Sim\}} -\log norm(f(t_j, s_i))$$

Quanto mais baixo for o valor de F, mais provável é que as palavras de um dado par sejam a tradução uma da outra.

O resultado é uma lista com as candidatas fonte e alvo, lado a lado, juntamente com a pontuação final de cada pareamento fonte e alvo. Essa pontuação final dá a possibilidade das candidatas serem classificadas em um *ranking* da mais provável para a menos

provável (na nossa pontuação, quanto menor o valor, maior é a probabilidade, explicaremos a razão disso acontecer no próximo capítulo quando explicamos os experimentos).

A tabela 4.4 traz exemplos de entradas nas listas fonte, alvo e bilíngues geradas por esse primeiro método em um de nossos experimentos.

Tabela 4.4: Exemplo de entradas presentes nas listas de candidatas a EM fonte, alvo e bilíngue geradas como saída do primeiro método

Candidatas fonte (português)	Candidatas alvo (inglês)	Final
realizar trabalho	carry work	0.98
	carry study	1.95
	carry research	2.45
	carry piece	2.98
	carry development	3.07
tomar lugar	take place	0.52

4.2 Segundo método de descoberta bilíngue

A Figura 4.2 ilustra a arquitetura proposta para o segundo método de descoberta bilíngue desenvolvido neste trabalho. Assim como na arquitetura apresentada para o primeiro método, os retângulos com cantos arredondados representam processos e os demais retângulos representam recursos.

Neste segundo método, a descoberta monolíngue ocorre apenas no lado fonte. A tradução não é realizada através da filtragem de duas listas de descoberta monolíngue, como na proposta anterior, mas sim gerando palavras similares alvo com base nos vetores bilíngues e córpus alvo.

A seguir descrevemos os passos usados para a descoberta bilíngue do segundo método:

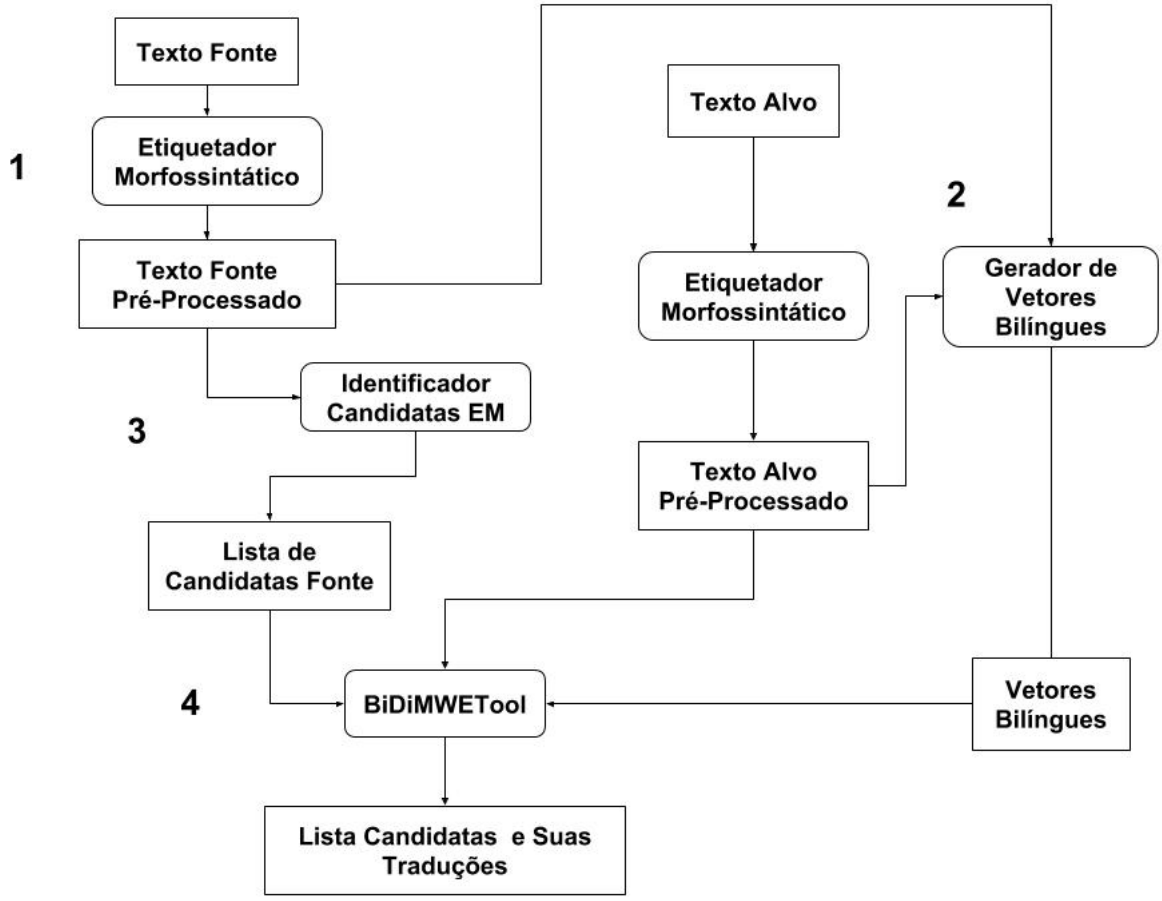
1. O córpus paralelo (da mesma maneira em que acontece no primeiro método) é fornecido como entrada. Esse córpus, contendo pares de texto fonte e texto alvo alinhados por sentença , etiquetado morfossintaticamente.

2. Novamente, neste passo os pares de textos paralelos são, também, usados pela ferramenta responsável por gerar vetores de palavras bilíngues.

3. Aqui também ocorre a descoberta monolíngue, mas diferentemente do primeiro método, a descoberta acontece apenas para a língua fonte.

4. Nesse segundo método implementado no *BiDiMWETool*, as possíveis traduções alvo, para cada candidata a EM fonte são recuperadas passando cada palavra da candidata

Figura 4.2: Arquitetura do segundo método de descoberta automática de EMs a partir de córpus paralelo no *BiDiMWETool*.



de forma separada na consulta feita aos vetores bilíngues. Por exemplo, se tivermos a candidata *realizar processo*, podemos recuperar as 15 palavras alvo mais similares para *realizar* (conforme Tabela 4.3) e as 15 palavras alvo mais similares para *processo*. Esse processo é exemplificado visualmente na Figura 4.3.

Em seguida, de forma similar ao que ocorre no primeiro método, é feito o pareamento das candidatas fonte e alvo com base nas sentenças (linhas) de ocorrência das candidatas. Para tanto, para cada sentença de ocorrência da candidata fonte, fazemos a intersecção das palavras que ocorrem na mesma sentença no córpus alvo e aquelas recuperadas pela consulta aos vetores bilíngues como possíveis traduções, como ilustra a figura 4.4. É importante mencionar que o córpus alvo está etiquetado morfossintaticamente.

Diferentemente do primeiro método, neste segundo, não foi necessário calcular nem a probabilidade (*P*) nem a similaridade (*Sim*) para definir o pareamento fonte-alvo. Desse modo, nenhuma pontuação final é gerada. Entretanto, durante a filtragem de candidatas, a comparação de similaridade entre candidatas fonte e alvo pode ser necessária, pois às

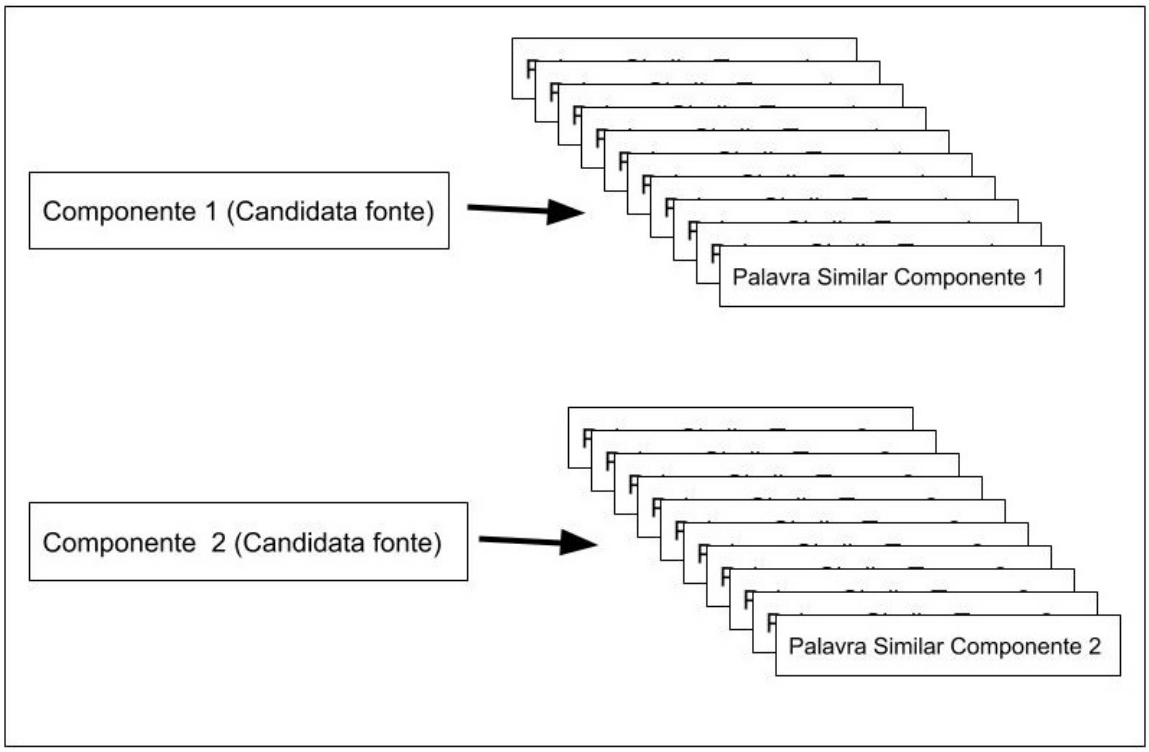


Figura 4.3: Geração de palavras similares alvo da candidata fonte: a candidata fonte é dividida em Componente 1 e Componente 2. Usando os vetores bilíngues, cada termo irá gerar palavras similares alvo.

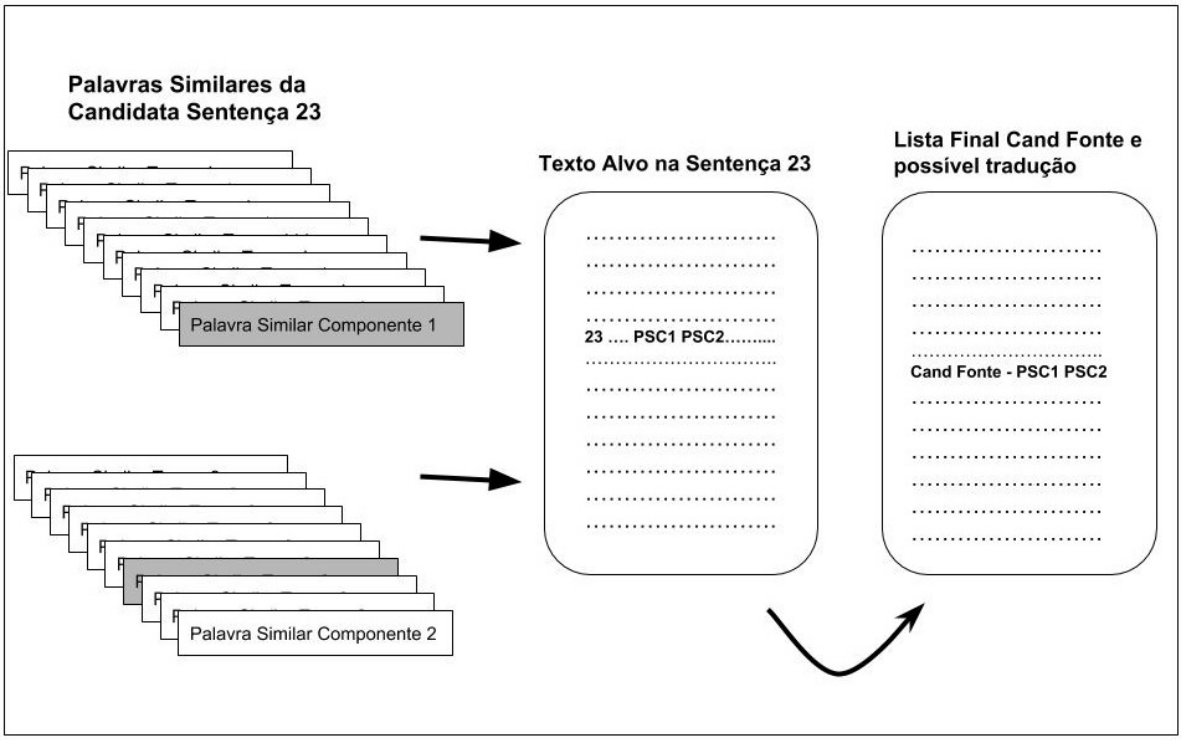


Figura 4.4: Cada número da sentença original fonte que possuía a candidata fonte foi guardado. Procura-se na mesma sentença de mesmo número do lado alvo as palavras similares geradas que ocorrem lá. Se houver palavras em comum, as palavras encontradas são retornadas como tradução da candidata fonte.

vezes, com a geração de candidatas alvo a partir de uma fonte, mais de duas palavras podem ser encontradas numa determinada linha, tornando necessário o uso de alguns filtros para eleger a melhor candidata. Tais filtros serão melhor explicadas no próximo capítulo.

A saída do *BiDiMWETool* para esse segundo método é a lista de candidatas da descoberta monolíngue fonte e a lista bilíngue. Para essa lista final de traduções, os pares de candidatas fonte e alvo que se repetem são agrupados e acompanhados da sua frequência no córpus. Esse método gera traduções juntamente com suas etiquetas morfossintáticas para fins de consulta. Isso será melhor explicado no capítulo 5.

Tabela 4.5: Exemplo de entradas presentes nas listas de candidatas a EM fonte e e alvo geradas como saída do segundo método

Candidatas fonte (português)	**Candidatas alvo (inglês)**	**Frequência**
tomar posse	V_take N_possession	3
tomar parte	V_take N_part	2
fazer viagem	V_do N_journey	2
	V_carry V_travel	1
	V_make N_journey	1
	V_make N_trip	1
	V_travel	1

4.3 Considerações sobre os métodos desenvolvidos

Os dois métodos de descoberta bilíngue foram desenvolvidos e testados tendo como base a premissa de realizar a descoberta bilíngue sem depender do alinhamento de palavras. A primeira proposta usa a ideia de juntar candidatas fonte e alvo descobertas pelo método bilíngue. Para que essa proposta funcione, é necessário fazer buscas de um lado fonte parecidas com o lado alvo. Um exemplo é quando procuramos no lado fonte pelo padrão *realizar* + *substantivo* e no lado alvo, pelo padrão *carry* + *substantivo*. Esse método, entretanto, tem a limitação de ser necessário definir um certo tipo de tradução, desde o início, definindo candidatas alvo como possíveis traduções já na busca monolíngue.

Já na segunda proposta, desenvolvemos um método que usa apenas a lista de candidatas monolíngue da língua fonte como entrada para a geração da lista bilíngue. Esse método tem a vantagem de ser menos dependente do conhecimento prévio da língua alvo. Por exemplo, digamos que temos um par de textos paralelos em português e alemão e não temos muito conhecimento de alemão. Com a segunda proposta é possível gerar traduções para a língua alvo não sendo necessário fazer a descoberta monolíngue do lado alvo.

Capítulo 5

EXPERIMENTOS E RESULTADOS

As duas primeiras seções deste capítulo apresentam os materiais usados nos experimentos aqui descritos: a seção 5.1 descreve o córpus usado, enquanto a seção 5.2 descreve as ferramentas utilizadas na implementação do *BiDiMWETool*. Os experimentos realizados serviram como base para validarmos os dois métodos apresentados até o momento. Por fim, são apresentados os resultados obtidos nos experimentos com os dois métodos desenvolvidos para a descoberta bilíngue no *BiDiMWETool*.

5.1 Córpus

O córpus FAPESP (AZIZ; SPECIA, 2011) foi usado para os experimentos. Esse córpus é composto por textos originalmente escritos em português do Brasil e traduzidos para sua versão em inglês contendo as quantidades de sentenças e *tokens* apresentadas na tabela 5.1.

Tabela 5.1: Número de sentenças e de *tokens* do córpus FAPESP

	Português	**Inglês**
sentenças	166.799	166.799
tokens	4.191.492	4.499.064

A Tabela 5.2 traz excertos de textos do córpus paralelo FAPESP com algumas sentenças fonte e alvo marcadas com candidatas a EM tanto do lado fonte quanto do lado alvo.

Tabela 5.2: Exemplos de sentenças paralelas do córpus FAPESP nas quais aparecem destacadas candidatas a EM fonte (português) e alvo (inglês)

Português	Inglês
Devemos **realizar** o primeiro **vôo** no final deste ano.	We expect to **conduct** the first **flight** this year.
Os pesquisadores também **realizaram projetos** piloto com algumas empresas dos setores agrícola e florestal para sondar a receptividade do mercado e avaliar a utilidade do sistema.	The researchers also **carried out** pilot **projects** with some companies from the agricultural and forestry sectors to sound out the receptivity of the market and to evaluate the usefulness of the system.
É importante conhecer todas as variáveis do HIV-1 para se **tomar medidas** certas de prevenção, diagnóstico e, especialmente, de tratamento.	It is important to know about all of the variations of HIV-1 so as to **take** the appropriate **measures** of prevention, diagnosis and especially of treatment.

5.2 Ferramentas

As próximas subseções descrevem as três ferramentas computacionais em uso neste trabalho: o etiquetador morfossintático TreeTagger[1] (SCHMID, 1994) , o *toolkit* para descoberta monolíngue de EMs MWEToolkit[2] (RAMISCH, 2015) e o gerador de vetores de palavras bilíngues MultiVec[3] (BÉRARD et al., 2016).

5.2.1 TreeTagger

Neste trabalho, o TreeTagger (SCHMID, 1994) foi a ferramenta adotada para essa tarefa, uma vez que aceita um córpus qualquer sendo capaz de processar diversas línguas tais como inglês, português, alemão, espanhol e russo. Após a etiquetação morfossintática dos córpus fonte (em português) e alvo (em inglês) usando o TreeTagger (etapa 1 dos métodos descritos nas Figuras 4.1 e 4.2), esses córpus etiquetados são utilizados na geração das listas de candidatas a EM pelo *MWEToolkit*.

5.2.2 MWEToolkit

O MWEToolkit (RAMISCH, 2015) foi a ferramenta escolhida para a etapa 3 dos métodos descritos nas Figuras 4.1 e 4.2. Esse *toolkit* agrupa *scripts* que realizam a descoberta automática de EMs de um córpus monolíngue. Além da descoberta automática realizada a partir de padrões morfossintáticos, os *scritps* do *toolkit* geram uma lista de candidatas ranqueada com base em

[1] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[2] http://mwetoolkit.sourceforge.net/PHITE.php
[3] https://github.com/eske/multivec

um conjunto de medidas de associações.

O *toolkit* foi desenvolvido como um conjunto de *scripts* em *python* e trabalha com arquivos XML que representam o córpus, os padrões a serem extraídos, a lista de candidatas e a saída. O *toolkit* também aceita outros formatos para córpus, como CoNLL[4], e pode produzir saídas em formato CSV. O fluxo de processamento do MWEToolkit é ilustrado na Figura 5.1, retirada do artigo original que descreve a ferramenta (RAMISCH; VILLAVICENCIO; BOITET, 2010).

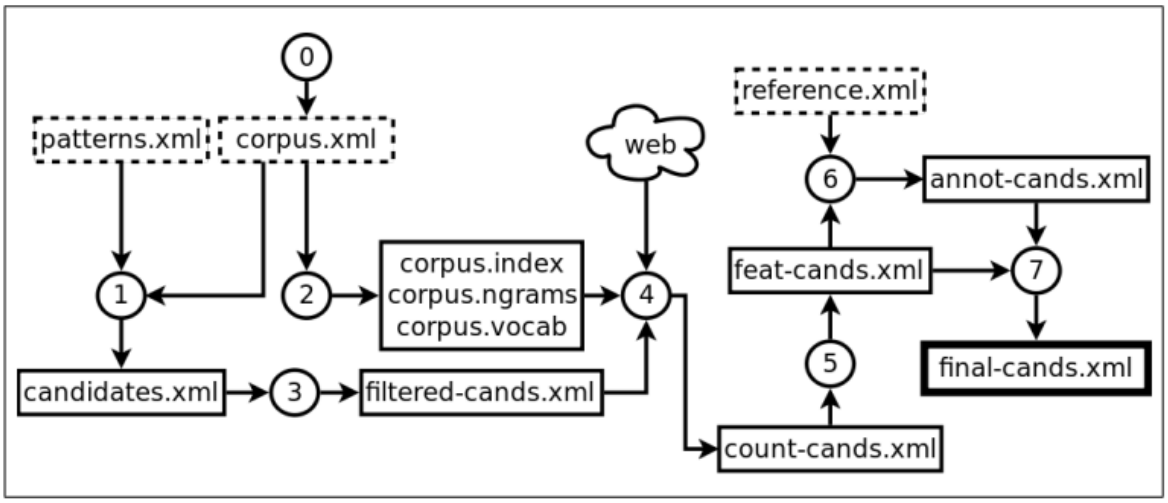


Figura 5.1: Arquitetura do *mwetoolkit*: no passo (0) o córpus é pré-processado, em seguida no passo (1) acontece a extração de candidatas que satisfazem os padrões morfossintáticos, (2) o córpus é indexado usando *arrays* de sufixo, (3) uma lista de candidatas é filtrada, (4) ocorre uma contagem dos n-gramas e palavras no córpus, (5) é feito o cálculo das medidas de associação e atributos descritivos, (6) é realizada uma anotação automática de (parte) das candidatas e (7) é realizado um treinamento/aplicação de um modelo de aprendizado de máquina. Entradas são mostradas como pontilhados, saídas são linhas mais grossas. Não fazemos uso dos passos 6 e 7 para a nossa extração monolíngue. Fonte: (RAMISCH; VILLAVICENCIO; BOITET, 2010)

Antes de rodarmos os *scripts* é aconselhável que o córpus de entrada seja pré-processado (passo 0 da Figura 5.1). Para isso é necessário rodar alguma ferramenta que possibilite o córpus ser etiquetado com informações morfossintáticas. Como explicado na seção anterior, essas informações consistem em dizer qual é a classe gramatical de cada palavra, seu lema e outros traços morfológicos. Esse córpus de entrada etiquetado (`corpus.xml`) é processado pelo *toolkit* que gera uma primeira lista de candidatas (`candidates.xml`) em n-gramas puros ou em padrões de etiquetas morfossintáticas (passo 1 da Figura 5.1).

Para otimizar o cálculo das medidas de associação explicadas na seção 2.2, o córpus é indexado usando um *array* de sufixo (passo 2 da Figura 5.1), que é uma estrutura de dados que faz uso eficiente da memória e permite que os n-gramas de tamanhos arbitrários sejam

[4]http://universaldependencies.org/format.html

encontrados em um córpus de tamanho grande.

Para cada sequência de n palavras contíguas p_1 até p_n, a ferramenta conta o número individual de ocorrências, $f(p_1)...f(p_n)$, e a frequência geral, $f(p_1....p_n)$ a partir do índice (`corpus.index`, `corpus.ngrams` e `corpus.vocab`).

O passo 3 da figura 5.1 gera uma lista de candidatas filtradas. Essa filtragem pode ser realizada para excluir candidatas que contenham pontuação, n-gramas que apareçam abaixo de algum limiar ou retirar algumas palavras específicas da lista (`filtered-cands.xml`). Após isso (passo 4), é calculada a frequência observada das candidatas e das palavras que as compõem, como explicado na seção 2.2 (`count-cands.xml`). Com esses valores (passo 5), são calculadas as medidas *Maximum likelihood estimator*, coeficiente *dice*, *pointwise mutual information*, *student's t-score* e log-likelihood (`feat-cands.xml`).

Se o usuário não tiver intenção de anotar as candidatas, é possível ir do passo 5 para o passo 7. Para a geração das candidatas do lado fonte e alvo do nosso trabalho, não usamos o recurso de anotação do MWEToolkit.

Ao final do processamento do *toolkit* tem-se uma lista de EMs (`final-cands.xml`). As listas contendo candidatas fonte e alvo (primeiro método) ou só a lista contendo candidatas fonte (segundo método) são/é fornecida(s) como entrada para o *BiDiMWETool*.

5.2.3 Multivec

Por fim, para a implementação da etapa 2 dos métodos descritos nas Figuras 4.1 e 4.2, escolheu-se o MultiVec (BÉRARD et al., 2016). Essa ferramenta gera e permite a manipulação de modelos de vetores de palavras como descrito na seção 2.4, em particular para modelos multilíngues. Além de combinar várias técnicas presentes na literatura, permite computar representações bilíngues a partir de córpus paralelos. Neste trabalho, o Multivec é usado para calcular similaridades entre duas línguas, como um auxílio ao processo de descoberta bilíngue de candidatas. A funcionalidade presente na ferramenta é:

1. Vetores Bilíngues de Palavras: para implementar essa funcionalidade, o Multivec usa o *bivec* proposto por Luong, Pham e Manning (2015) que, por sua vez, possui os mesmos métodos do *word2vec*, porém realiza o treinamento em cima de um córpus paralelo gerando vetores de palavras de forma bilíngue.

   Para cada par de sentenças no córpus paralelo, o bivec tenta predizer as palavras da mesma forma que o *word2ec* faz, mas também usa as palavras da sentença fonte para predizer as

Tabela 5.3: Padrões morfossintáticos aplicados nos córpus em português e em inglês

	Padrões para o português	Padrões para o inglês
1	Ficar + ADJETIVO	[*Get* | *Become* | *Be*] + ADJETIVO
2	Realizar + SUBSTANTIVO	[*Carry* | *Make*] + SUBSTANTIVO
3	[Fazer | Dar | Tomar] + SUBSTANTIVO	[*Make* | *Take* | *Give*] + SUBSTANTIVO

palavras da sentença alvo e vice e versa. Portanto, para cada atualização de pesos feita no *word2vec*, o *bivec* realiza 4 atualizações: fonte para fonte, fonte para alo, alvo para alvo, e alvo para fonte.

Os vetores bilíngues, gerados pelo MultiVec, também são fornecidos como entrada para o *BiDiMWETool*.

5.3 Experimentos

Essa seção descreve dois experimentos realizados para avaliar as duas propostas para descoberta automática de EMs em córpus paralelo. No primeiro experimento, apresentado na subseção 5.3.1, gerou-se uma lista de referência de EMs em português e suas traduções para o inglês. No segundo experimento, descrito na subseção 5.3.2, avaliou-se a efetividade do uso de vetores bilíngues de palavras para a descoberta de EMs e suas traduções, a partir de córpus paralelo.

5.3.1 Experimentos com o primeiro método de descoberta bilíngue

Esta seção explica as hipóteses levantadas e os experimentos realizados a fim de conceber a abordagem do primeiro método de descoberta bilíngue abordado na seção 4.1 do capítulo 4. A ideia geral desse primeiro experimento era gerar uma lista de referência de EMs em português e suas correspondentes traduções para inglês. Para tanto, adotou-se como hipótese o fato de que algumas categorias de expressões verbais em português são traduzidas para expressões verbais em inglês. O córpus utilizado foi o FAPESP (AZIZ; SPECIA, 2011) e decidiu-se pela extração de construções verbo suporte do tipo *realizar pesquisa, tomar banho* e *tirar foto*, explicadas no Capítulo 2.

Para esse primeiro experimento, seis passos foram realizados, como explicado a seguir:

1. Definição dos padrões morfossintáticos: padrões morfossintáticos foram definidos para extrair as candidatas tanto no córpus fonte (português) quanto no córpus alvo (inglês).

Os padrões para a extração no português foram definidos de acordo com a observação de ocorrência no córpus; já os padrões para o inglês foram definidos por meio da tradução dos padrões fonte. A Tabela 5.3 traz todos os padrões usados para ambos os idiomas. Os padrões morfossintáticos permitiam que o verbo seguido do complemento fossem não contíguos, ou seja, era permitido que outras palavras ocorressem no meio, como mostrado na tabela 5.2

2. Extração monolíngue de candidatas: usando o MWEToolkit, realizamos a extração das candidatas e geramos duas listas com os resultados: uma para o português e outra para o inglês. Cada candidata é acompanhada do número de ocorrências (frequência no córpus), em qual(is) sentença(s) do córpus ocorreu e medidas de associação que mostram as chances dela ser considerada, de fato, uma EM.

3. Pareamento das candidatas fonte e alvo: o próximo passo foi relacionar ambas as listas no intuito de encontrar as candidatas alvo que fossem possíveis traduções das candidatas na lista fonte. Esse pareamento foi realizado considerando-se as sentenças (linhas) de ocorrência para as candidatas em português e inglês. Por exemplo, se uma candidata apareceu na linha 5 do córpus alvo, buscou-se as candidatas do córpus fonte que apareceam na linha 5 também. Essa estratégia se baseou no fato de que os córpus em português e em inglês são a tradução um do outro e, desse modo, cada linha em português é traduzida (está alinhada) com a linha correspondente em inglês.

4. Cálculo da probabilidade condicional (P): calculamos a probabilidade das candidatas como explicado na seção 4.1.

5. Cálculo da similaridade (*Sim*): calculamos a similaridade distributiva entre os pares de candidatas fonte e alvo, conforme descrito na seção 4.1.

6. Cálculo da pontuação final (F): a pontuação final de um par de candidatas fonte-alvo é calculada com base em três medidas: P, t-score (veja seção 2.2) da candidata alvo e *Sim*, conforme descrito na seção 4.1.

Nas Tabelas 5.4, 5.5 e 5.6 pode-se ver uma amostra dos resultados obtidos neste primeiro experimento para cada um dos três padrões definidos na Tabela 5.3. As tabelas mostram as candidatas fonte (EM fonte), seguidas de suas possíveis traduções (EM alvo), juntamente com o número de vezes que as expressões aparecem na mesma linha nos textos paralelos (# par), a medida *t-score* calculada para a EM alvo (t-score alvo), o grau de similaridade entre elas (*Sim*) e, por último, a pontuação final (F). Nessas tabelas, os pares de candidatas (EM fonte, EM

alvo) estão ordenados crescentemente pelo valor de F, lembrando que quanto menor esse valor, mais provável é de o par estar correto. Além disso, destaca-se em negrito os pares considerados corretos.

Tabela 5.4: Resultados para o padrão ficar+ADJETIVO e [*get* | *become* | *be*] + ADJETIVO

	EM fonte	EM alvo	# par	*t*-score alvo	Sim	F
1	**ficar doente**	**get sick**	2	0.51	0.53	1.44
2	**ficar doente**	**become ill**	2	0.50	0.46	1.51
3	ficar doente	be normal	1	0.52	0.41	1.84
4	**ficar doente**	**become sick**	1	0.49	0.41	1.86
5	ficar doente	be tolerant	1	0.50	0.33	1.95
1	**ficar pronto**	**be ready**	46	0.72	0.67	0.41
2	**ficar pronto**	**become ready**	5	0.50	0.60	1.58
3	**ficar pronto**	**get ready**	1	0.58	0.69	2.15
4	ficar pronto	be capable	2	0.74	0.25	2.18
5	ficar pronto	be necessary	1	0.87	0.40	2.22
6	ficar pronto	be fundamental	1	0.74	0.26	2.47

Tabela 5.5: Resultados para o padrão realizar + SUBSTANTIVO e [*carry* | *make*] + SUBSTANTIVO.

	EM fonte	EM alvo	# par	*t*-score alvo	Sim	F
1	**realizar teste**	**carry test**	20	0.50	0.73	0.71
2	**realizar teste**	**carry trial**	3	0.29	0.63	1.85
3	realizar teste	carry field	4	0.26	0.47	1.89
4	realizar teste	make assessment	4	0.22	0.28	2.18
5	realizar teste	make use	1	1.00	0.24	2.19
6	**realizar teste**	**make test**	1	0.23	0.62	2.43
7	realizar teste	make comparison	1	0.38	0.30	2.51
8	**realizar teste**	**carry test**	1	0.17	0.65	2.54
9	realizar teste	carry safety	1	0.17	0.53	2.64
10	realizar teste	make prototype	1	0.23	0.37	2.64
11	realizar teste	make search	1	0.15	0.24	3.02
1	realizar substituição	carry identification	1	0.19	0.45	1.67

Com relação ao padrão *carry* + SUBSTANTIVO, vale mencionar que a, que aparece na Tabela 5.5, partícula *out*, que seria o complemento de *carry*, foi suprimida da saída final para facilitar o cálculo.

Para a Tabela 5.6, verificamos que as últimas comparações possuem valores de similaridade (*Sim*) muito baixos. Nesse caso, a expressão em português possui grandes chances de ser traduzida como uma só palavra no inglês. Esses valores de similaridade baixos ocorrem quando o Multivec não encontra palavras do lado fonte e alvo ocorrendo no mesmo contexto com frequência. Por exemplo, certas palavras como *fazer* são traduzidas como *make*. Se quisermos comparar a similaridade entre *fazer* e a palavra *walk*, vamos obter um valor de similaridade

Tabela 5.6: Resultados para o padrão [fazer | dar] + SUBSTANTIVO e [*make* | *do*] + SUBSTANTIVO

	EM fonte	EM alvo	# par	*t*-score alvo	Sim	F
1	**fazer comparação**	**make comparison**	4	0.37	0.64	1.16
2	**fazer comparação**	**do comparison**	1	0.23	0.56	2.04
3	fazer comparação	make method	1	0.21	0.44	2.18
4	fazer comparação	make drug	1	0.23	0.33	2.27
1	dar início	do thing	4	0.44	0.15	1.76
2	dar início	do Sul	1	1.00	0.13	2.06
3	dar início	make vaccine	1	0.31	0.24	2.30
4	dar início	make list	1	0.26	0.24	2.37
5	dar início	make roster	1	0.24	0.21	2.47

Tabela 5.7: Novos Padrões morfossintáticos aplicados nos córpus em português e em inglês

	Padrões para o português	Padrões para o inglês
1	[Entrar| Colocar] + VERBO	VERBO + SUBSTANTIVO
2	Realizar + SUBSTANTIVO	[*Carry* | *Make*] + SUBSTANTIVO
3	Dar + SUBSTANTIVO	VERBO + SUBSTANTIVO
4	[Fazer | Tomar] + SUBSTANTIVO	[*Make* | *Take*] + SUBSTANTIVO

muito baixo, pois o método irá verificar que *fazer* no lado fonte raramente co-ocorre com a palavra do lado alvo.

Após os primeiros experimentos, comparando os resultados da extração monolíngue, com a lista final de traduções, verificamos que em certos casos, as traduções para alguns verbos não eram exatamente as esperadas. Por exemplo, a expressão *dar ênfase* pode ser traduzida como *place emphasis*, *lay emphasis* e *put emphasis*. Nota-se que se definirmos como *give* o padrão do lavo alvo, o método poderia deixar outras traduções de fora. Deste modo, definir o que estivéssemos esperando do lado alvo poderia deixar muitas traduções de fora. Por esse motivo, para alguns padrões em inglês, decidimos deixar o padrão alvo mais "relaxado", isto é, fazíamos uma busca apenas por qualquer verbo que fosse seguido de um substantivo (de forma não contígua como antes). Infelizmente, quando aplicávamos esse padrão, o número de comparações na lista final aumentava, pois por exemplo, em uma linha o padrão *verbo + substantivo* poderia aparecer três vezes em uma linha e ser comparado todas as vezes com o padrão fonte. A tabela 5.7 mostra algumas modificações realizadas para definir novos padrões morfossintáticos aplicados ao método. Os verbos *colocar* e *entrar* foram observados no córpus e passaram a fazer parte de um dos padrões fonte. Decidimos retirar o verbo *dar* e *give* dos padrões fonte e alvo que aparecem no padrão da linha 4 da tabela 5.5.

Por exemplo, na tabela 5.8, para o padrão *dar sentido*, verificamos que ao relaxarmos a busca monolíngue alvo para qualquer verbo e o substantivo mais próximo, o nosso método

Tabela 5.8: Resultados para o padrão dar + SUBSTANTIVO e VERBO + SUBSTANTIVO.

	EM fonte	EM alvo	# par	*t*-score alvo	Sim	F
1	**dar sentido**	**provide meaning**	3	0.973	0.59	1.47
2	**dar sentido**	**make sense**	3	0.983	0.57	1.48
3	**dar sentido**	**give sense**	1	0.972	0.75	1.86
4	**dar sentido**	**give meaning**	1	0.973	0.65	1.93
5	dar sentido	give side	1	0.972	0.48	2.06
6	dar sentido	provide purpose	1	0.972	0.42	2.12
7	dar sentido	save Brazil	1	0.972	0.39	2.15
8	dar sentido	take part	1	0.999	0.35	2.19
9	dar sentido	understand everything	1	0.973	0.31	2.25
10	dar sentido	be party	1	0.972	0.31	2.25
1	**dar ênfase**	**place emphasis**	2	0.973	0.53	1.71
2	**dar ênfase**	**lay emphasis**	1	0.973	0.60	1.98
3	dar ênfase	guarantee quality	1	0.976	0.43	2.12
4	dar ênfase	gain status	1	0.975	0.42	2.13
5	**dar ênfase**	**emphasize** issue	1	0.972	0.39	2.16
6	dar ênfase	encourage reading	1	0.973	0.39	2.17
7	**dar ênfase**	**emphasize** teaching	1	0.972	0.35	2.21

conseguiu fazer um pareamento em que os verbos de tradução fossem *provide* e *make*, uma vez que a tradução literal de dar é *give*. Entretanto, com a abordagem de "relaxar" o lado alvo, pode haver muitos pareamentos como nos casos da mesma tabela, onde candidatas como *save Brazil* foram pareadas por terem sido recuperadas na busca alvo. Outro ponto a ser levantado é para a candidata *dar ênfase*, nas linhas 5 e 7, onde podemos verificar que a tradução seria apenas *emphasize*. Ou seja, nesses casos, a tradução consistia de apenas um verbo. Como a busca monolíngue alvo foi projetada, no primeiro método, para retornar sempre um *verbo + substantivo*, o método pareou duas palavras ao invés de uma. Por esse motivo, pensamos em um outro método em que não usássemos a extração monolíngue do lado alvo e, usaríamos os vetores bilíngues como um "gerador" de traduções, de acordo com o espaço vetorial. A próxima subseção aborda e mostra os resultados obtidos com esse novo método. A avaliação final deste primeiro método será mostrado na seção 5.4.

5.3.2 Experimentos com o segundo método de descoberta bilíngue

Para o segundo experimento, usamos exclusivamente os vetores de palavras para tentar mapear as possíveis traduções de candidatas. Outra diferença em relação ao primeiro experimento é que, agora, fazemos a extração monolíngue apenas no córpus fonte, pois esperamos que as traduções sejam obtidas usando os vetores gerados pelo MultiVec.

Assim, os passos para esse segundo experimento foram:

1. Definição dos padrões morfossintáticos: foram usados os mesmos padrões fonte definidos para o primeiro experimento (veja Tabela 5.3).

2. Extração monolíngue de candidatas: novamente, o MWEToolkit foi usado para extrair as candidatas, como no primeiro experimento. Desta vez, contudo, apenas candidatas fonte foram extraídas.

3. Busca por possíveis traduções: as possíveis traduções alvo, para cada candidata a EM fonte, foram recuperadas passando cada palavra da candidata de forma separada para o Multivec. Este passo corresponde ao passo 4 mostrado na seção 4.2 e tomamos como 10 o número de palavras similares a serem geradas pelos vetores bilíngues.

4. Pareamento das candidatas fonte e alvo: Por fim, realiza-se o pareamento das candidatas conforme descrito na seção 4.2. Nesse primeiro momento dos experimentos, como geramos apenas 10 palavras para cada termo da candidata fonte, não foi necessário aplicar filtros pois quando geramos poucas candidatas alvo através dos vetores bilíngues, não temos muitas palavras para procurar no texto alvo.

A Tabela 5.9 traz o resultado preliminar deste segundo experimento. Para fins de visualização, apresentamos o lema de cada palavra da candidata. Contudo, nossos primeiros experimentos usamos as formas superficiais das palavras para as buscas com o MultiVec. Com base nesses resultados preliminares nota-se que é possível encontrar boas candidatas a tradução para uma EM fonte usando apenas os vetores bilíngues de palavras gerados pelo MultiVec. Nos próximos experimentos usamos o lema das palavras para gerar os vetores bilíngues e para consulta no córpus alvo, para evitar que várias formas da mesma palavra apareçam nas traduções, por exemplo *carry — carried — carrying out tests*.

A partir dos exemplos da Tabela 5.9 também constata-se que, em alguns casos, a tradução encontrada é apenas uma palavra (o verbo) e não uma expressão multipalavra, como podemos ver na linha 17 da tabela, e da mesma forma que já tínhamos observado durante os experimentos da primeira proposta. Os dados foram ordenados de acordo com o número de ocorrências ($\#T$). Todas as vezes que uma candidata fonte teve uma tradução específica, por exemplo, *realizar testes* foi traduzida como *carried out tests* 15 vezes. Cabe mencionar que esses casos eram um problema a ser resolvido no método anterior, uma vez que ele era incapaz de encontrar traduções corretas entre uma EM fonte e uma única palavra alvo devido à estratégia adotada que extraía candidatas tanto no córpus fonte quanto no córpus alvo e depois encontrava suas correspondências.

Tabela 5.9: Traduções encontradas para a candidata Realizar + Teste. Todas as ocorrências mostram-se aptas a serem consideradas traduções da candidata fonte.

	Cand fonte	Tradução encontrada	# T
1	realizar teste	carried out tests	15
2	realizar teste	carry out tests	10
3	realizar teste	carrying out tests	5
4	realizar teste	tests	2
5	realizar teste	carrying out trials	2
6	realizar teste	conduct tests	2
7	realizar teste	conducted tests	2
8	realizar teste	conducted trials	2
9	realizar teste	carried out test	1
10	realizar teste	carried out trials	1
11	realizar teste	conducting tests	1
12	realizar teste	conduct trials	1
13	realizar teste	conducting trials	2
14	realizar teste	carry out testing	1
15	realizar teste	carry out test	1
16	realizar teste	holding out trials	1
17	realizar teste	test	1

A partir desses resultados preliminares, contudo, ainda não era possível afirmar que o método era capaz de recuperar a tradução correta. Por exemplo, usando o mesmo exemplo da tabela 5.9, na linha 17, onde *realizar teste* foi traduzido como *test*, não podemos afirmar se o método realmente descobriu o verbo *test*, ou se por algum motivo o método apenas conseguiu recuperar o substantivo *test* na sentença que coincidentemente em inglês é idêntico ao verbo. Partindo do pressuposto que não estávamos considerando as classes gramaticais das palavras, decidimos realizar novamente o experimento levando em conta as informações morfossintáticas do córpus. Além do uso dos vetores bilíngues para gerarmos possíveis traduções, fizemos o uso do córpus alvo etiquetado morfossintaticamente para verificar se o nosso método estava recuperando as traduções com as classes gramaticais esperadas. As mudanças no método e os novos passos são:

1. Aplicamos os passos 1 e 2 do experimento anterior.

2. Busca por possíveis traduções: continuamos usando a geração de palavras similares através dos vetores bilíngues gerados pelo Multivec. Da mesma forma, para cada palavra da candidata fonte, usamos o Multivec para recuperar as 20 palavras mais similares. Decidimos aumentar o número de palavras similares geradas pois teríamos mais chances das *embeddings* retornarem palavras traduzidas com pouca frequência. Por exemplo, para a candidata *realizar pesquisa* o método gerou as 20 palavras em inglês mais pa-

recidas semanticamente com *realizar* e as 20 palavras mais parecidas semanticamente com *pesquisa*. Ao todo, até 40 palavras são recuperadas como possíveis traduções para a candidata em questão.

3. Pareamento das candidatas fonte e alvo: o próximo passo era localizar a sentença alvo do texto. Por exemplo, se *realizar pesquisa* aparecesse na linha 23 do lado fonte, procurávamos na linha 23 do texto alvo quais as palavras da nossa lista de 40 que aparecessem na sentença. O número de palavras gerado não precisa ser fixo. A ideia é que quanto mais palavras similares geramos, maiores são as chances de alguma delas de estar na sentença alvo. Quanto mais palavras geramos, mais palavras irrelevantes o método proposto pode retornar. O próximo passo é aplicar um filtro para selecionar quais são as palavras que serão selecionadas como tradução final do método.

4. Filtro das Candidatas Alvo: o filtro irá ser usado dependendo de quantas candidatas alvo o método retornou, para realizá-lo é necessário alguns passos e são eles:

- Se o número de palavras retornadas pelo método era maior que um:

    1. Realizamos o cálculo de *similaridade distributiva* entre cada palavra da candidata fonte e da candidata alvo e elegíamos os pares com as pontuações mais altas. Além disso, era necessário verificar se as palavras elegidas pela filtragem estavam na ordem original de ocorrência na frase. Por exemplo, se das 40 palavras que os vetores bilíngue geraram para uma candidata fonte (20 mais similares ao primeiro termo e 20, ao segundo), 5 fossem encontradas na sentença alvo essas 5 palavras eram comparadas com o primeiro termo da candidata fonte através da pontuação de similaridade. A palavra alvo mais similar era recuperada primeiro. As 4 palavras alvo restantes eram comparadas com o segundo termo e a mais similar era recuperada. Essas duas palavras alvo recuperadas eram consideradas como a candidata à tradução.

    2. O próximo passo era verificar a classe gramatical das palavras presentes na candidata à tradução. A ideia da descoberta era considerar sempre traduções do tipo VERBO + SUBSTANTIVO. Assim, traduções do tipo SUBSTANTIVO + SUBSTANTIVO eram eliminadas. Por exemplo, considere que o método tenha retornado *N_research N_survey* como candidata à tradução da candidata fonte *fazer pesquisa* ao invés de *perform survey*. A solução implementada para esse caso se baseou na busca pelo verbo necessário para compor a tradução considerando-se o número da sentença em que a candidata fonte ocorria. Uma vez que o índice da sentença alvo

tenha sido detectado, começamos uma busca nela pela segunda palavra, no caso do nosso exemplo é *N_survey*. Uma vez que essa palavra na sentença alvo era encontrada, procurávamos na sentença a palavra classificada como verbo mais próxima dela (a sua esquerda), e então esse verbo era colocado no lugar da primeira palavra originalmente encontrada. Para o nosso exemplo, uma vez que encontrássemos a palavra *N_survey*, fazíamos uma busca à esquerda da palavra, e o primeiro verbo encontrado retornado (*V_perform*) era colocado colocado no lugar da primeira palavra originalmente encontrada (*N_research*). Após a aplicação do nosso filtro, a tradução alvo final era *perform survey* e não *research survey*. Se por exemplo, usando esse método de filtro, não conseguíssemos identificar um verbo mais próximo (à esquerda), o método apenas retornaria o que ele já tinha gerado até o momento (a tradução final após o filtro continuaria sendo *N_research N_survey*). Esse filtro foi pensado para tratar as candidatas dos experimentos deste trabalho e funciona apenas para casos específicos como o caso de Construções Verbo Suporte com objeto direto. Para os demais casos esse filtro deve ser adaptado.

- Se o método retornava apenas uma palavra:

  1. Se essa palavra era classificada como um verbo, essa palavra era considerada a tradução final da candidata fonte. Verificamos que, de acordo com o primeiro experimento, muitas construções do tipo verbo suporte em português eram, na verdade, traduzidas apenas como um verbo simples em inglês.
  2. Se a palavra não era classificada como um verbo, realizávamos a busca pelo verbo mais próximo da mesma maneira que no caso anterior. Por exemplo, para a candidata fonte *realizar sonho* o método sem filtros encontrou apenas *dream*. Contudo, ao analisar também as classes gramaticais verificamos que essa palavra era um substantivo. Ao aplicarmos o método de procurar pelo verbo mais próximo, verificamos que o verbo em questão era o *fulfill*, ou seja, a candidata a tradução correta seria *fulfill dream*.

- Se o método não retornava nenhuma palavra, nenhuma candidata à tradução era gerada, indicando que o método não foi capaz de inferir uma possível tradução para a candidata fonte usando apenas os vetores bilíngues.

Vale mencionar que o método pode gerar algumas falhas na hora de filtrar substantivos por exemplo, uma vez que os etiquetadores morfossintáticos podem marcar palavras com classes gramaticais incorretas. Em alguns momentos, um substantivo era considerado verbo e poderia

representar um problema para a nossa descoberta. Entretanto, mesmo que casos em que classes gramaticais trocadas ocorressem no córpus etiquetado, os casos em que as classes gramaticais estavam corretas deram evidência de que traduções estavam corretas.

O resultado desse segundo método com o passo extra de filtragem baseada em etiquetas morfossintáticas também foi uma lista de candidatas fonte com suas traduções geradas pelos vetores bilíngues. Na tabela 5.10 podemos ver alguns resultados em que as traduções geradas estavam em alguma das formas esperadas: *verbo + substantivo* ou *verbo*. No lado esquerdo temos a candidata fonte e do lado direito sua tradução correspondente. O método também informou quantas vezes aquela candidata foi traduzida daquela forma. Por exemplo, *realizar projeto* foi traduzida como *carry project* sete vezes e como *create project* apenas uma vez. O método retornou a classe gramatical das palavras geradas pelo etiquetador morfossintático.

Tabela 5.10: Traduções encontradas para o padrão Realizar + Substantivo sem casos problemáticos

	EM fonte	**Tradução encontrada**	**# T**
1	realizar projeto	V_carry *(out)* N_project	7
2	realizar projeto	V_create N_project	1
1	realizar processo	V_carry *(out)* N_process	2
2	realizar processo	V_conduct N_process	1
1	realizar estudo	V_carry *(out)* N_study	28
2	realizar estudo	V_conduct N_study	8
3	realizar estudo	V_undertake N_study	1
4	realizar estudo	V_complete N_study	1
5	realizar estudo	V_make N_study	1
6	realizar estudo	V_study	1

A tabela 5.11 apresenta alguns exemplos para os quais o método não conseguiu recuperar algumas traduções da forma esperada. Para o caso em que a candidata *dar palestra* teve sua tradução recuperada como apenas *give*, verificamos que não é o mesmo caso de *realizar estudo* como *study*. Isso aconteceu porque, para este caso, a ocorrência de *dar palestra* no córpus foi traduzida como *give talk*, mas não de forma frequente. Essa baixa ocorrência de traduções fez com que a palavra *talk* não fosse recuperada pelos vetores bilíngues como uma possível tradução de *palestra* e, dessa forma, o método não procurou pela palavra *talk* no córpus alvo. Para os casos das traduções de *dar forma*, verificamos que não há traduções na candidata número 3 pois ao gerar as palavras similares pelos vetores bilíngues, nenhuma delas aparecia na sentença alvo correspondente.

A seguir apresentamos os contextos originais em que um desses casos (do método não inferir nenhuma tradução) ocorreram no córpus, mostrando a sentença fonte e alvo:

1) “Imagens no infravermelho desse lado B da tela revelaram uma terceira pintura escondida sob as tintas que **deram forma** à figura feminina (...)”

2) “Infrared images of this B side of the canvas revealed a third picture hidden under the paint that **was used to portray** the woman (...)”

Verificamos que nesse caso a tradução não foi recuperada por se tratar de uma tradução muito pontual e não frequente no texto, pois temos um padrão *dar forma* correspondendo ao verbo *portray*.

Tabela 5.11: Traduções encontradas para o padrão Dar + Substantivo com casos problemáticos

	EM fonte	Tradução encontrada	# T
1	dar palestra	V_give N_lecture	3
2	dar palestra	V_lecture	1
3	dar palestra	V_give	1
1	dar forma	V_shape	4
2	dar forma	V_give N_shape	3
3	dar forma	-	2
4	dar forma	V_give N_form	2
5	dar forma	V_form	1
6	dar forma	V_set N_form	1

5.4 Resultados e Avaliação dos Experimentos

Dados os resultados gerados pelos dois métodos na forma de lista, a avaliação dos pares (candidata a EM fonte, candidata a tradução) gerados pelos dois métodos foi feita manualmente por dois juízes humanos. Para ser um juiz nessa tarefa de avaliação foi necessário apenas ter um bom domínio da língua portuguesa. Eles analisaram um subconjunto dos pares gerados avaliando cada par da seguinte maneira:

EM fonte Se a candidata fonte (português) era considerada de fato uma EM (VP, Verdadeira Positiva termo mencionado no capítulo 2) ou não.

Tradução Se a candidata em português fosse considerada VP, analisava-se se a candidata alvo (inglês) pareada correspondia a uma tradução de forma inteira, de forma parcial, ou se não era uma possível tradução. Para essa tarefa, além das candidatas, os juízes humanos tiveram acesso às sentenças inteiras do córpus (fonte e alvo) onde as candidatas tinham ocorrido para fins de consulta.

Para realizar a análise das candidatas, usamos as diretrizes do PARSEME[5] e pode ser conferida também em Savary et al. (2017).

A Tabela 5.12 traz exemplos de sentenças paralelas nas quais é possível notar, em negrito, a ocorrência de candidatas a EM acompanhadas de suas possíveis traduções. Para cada par apresenta-se, também, a avaliação dada pelos juízes humanos: S (sim), N (não) ou P (para tradução parcialmente correta).

Tabela 5.12: Exemplos de sentenças paralelas com a ocorrência de candidatas a EM, acompanhadas de suas possíveis traduções e da avaliação dada pelos juízes

Sentença em português	**Sentença em inglês**	**EM?**	**Trad?**
LinkGen é a primeira empresa na América Latina a **fazer teste** de DNA para cavalos e bois.	LinkGen is the first company in Latin America to **do** a DNA **test** for horses and cattle.	S	S
Para **realizar** essa **modificação**, os pesquisadores escolheram e adicionaram ao processo dois catalisadores, que são substâncias promotoras da reação.	To achieve this, the researchers selected and **added** to the **process** two catalysts, i.e., substances that promote the reaction.	S	N
Além de afetar o número de estômatos, a mutação Well produz estruturas anatômicas que permitem que a planta **realize** mais **fotossíntese** com menos água.	In addition to affecting the number of stomata, the Well mutation produces anatomic structures that **allow** the plant to produce more **photosynthesis** with less water.	S	P

Na descoberta bilíngue realizada pelos dois métodos, pareamos as candidatas obtidas pela extração monolíngue fonte (português) e as candidatas obtidas pela extração monolíngue alvo (inglês). O pareamento se baseia nas linhas de ocorrência das candidatas fonte (no córpus fonte) com as linhas de ocorrência das candidatas alvo (no córpus alvo). Para realizarmos o cálculo de precisão e MAP do nosso experimento, analisamos os pares na medida que eles apareciam sem considerarmos sua repetição na lista final. Vejamos o exemplo na Tabela 5.13, onde a candidata *Dar Palestra* é pareada com *use book*, *give talk* e *give lecture*. Notamos que os dois primeiros pares aparecem apenas uma vez, mas o terceiro par aparece três vezes. Notamos também que o primeiro par não é considerado PV mas os outros dois sim. Mesmo que o último par tenha aparecido 3 vezes na nossa lista final, vamos considerar essa ocorrência como apenas um acerto, e não três. No exemplo então, teríamos um acerto de 2 para 3 ocorrências, e não 4 acertos para 5 ocorrências.

Para a seleção do subconjunto de pares a serem avaliados, usamos o *ranking* da lista de candidatas fonte com base na medida *T-Student Score*. As candidatas fonte com maior pontuação

[5]http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/

Tabela 5.13: Pareamento do padrão Dar + SUBSTANTIVO e Verbo + SUBSTANTIVO.

Candidata fonte	Candidata alvo	Frequência do par	EM?
dar palestra	use book	1	Não
dar palestra	give talk	1	VP
dar palestra	give lecture	3	VP

foram elegidas para serem analisadas juntamente com suas possíveis traduções.

5.4.1 Resultados dos experimentos com o primeiro método de descoberta bilíngue

Na tabela 5.14 estão listados o número de candidatas retornadas para cada padrão fonte (em português) e, na tabela 5.15, para cada padrão alvo (em inglês) no primeiro experimento. A primeira coluna mostra os padrões usados, a segunda coluna apresenta o número de candidatas sem repetição, em outras palavras, a quantidade de padrões diferentes e, por último, as candidatas retornadas ao todo.[6] Por exemplo, o padrão *Dar + Substativo* gerou 374 candidatas diferentes em nosso córpus (*dar início*, *dar origem* ...). Entretanto, como algumas dessas candidatas pertencentes a esse padrão ocorreram mais de uma vez ao longo do córpus, no total foram recuperadas 1472 candidatas.

Tabela 5.14: Número de candidatas retornadas após a aplicação dos padrões morfossintáticos no córpus em português

	Padrões em português	# candidatas diferentes	# total de candidatas
1	Realizar + SUBSTANTIVO	200	578
2	[Fazer | Tomar] + SUBSTANTIVO	807	2975
3	[Entrar| Colocar] + SUBSTANTIVO	499	1177
4	Dar + SUBSTANTIVO	374	1472
5	Ficar + ADJETIVO	143	488

Tabela 5.15: Número de candidatas retornadas após a aplicação dos padrões morfossintáticos no córpus em inglês

	Padrões em inglês	# candidatas diferentes	# total de candidatas
1	[*Carry* | *Make*] + SUBSTANTIVO	1840	5560
2	[*Make* | *Take* | *Do*] + SUBSTANTIVO	4277	14607
3	VERBO + SUBSTANTIVO	94061	193816
4	[*Get* | *Become* | *Be*] + ADJETIVO	2573	23879

Na tabela 5.16 mostramos o número de pares retornados após a aplicação do primeiro método juntamente com o número de pares que elegemos para analisar manualmente.

[6] O símbolo # denota a quantidade.

Tabela 5.16: Número de pares gerados pelo primeiro método de descoberta bilíngue

	Padrões em português	Padrões em inglês	# pares	# pares avaliados
1	Realizar + SUBSTANTIVO	[*Carry* | *Make*] + SUBSTANTIVO	155	99
2	[Fazer | Tomar] + SUBSTANTIVO	[*Make* | *Take*] + SUBSTANTIVO	903	97
3	[Entrar| Colocar] + VERBO	VERBO + SUBSTANTIVO	3224	213
4	Dar + SUBSTANTIVO	VERBO + SUBSTANTIVO	2468	99

A tabela 5.17 traz os resultados: a primeira coluna traz o número correspondente ao par de padrões (veja tabela 5.16), a segunda coluna mostra quantas candidatas foram consideradas VP do total de candidatas analisadas e as demais mostram o resultado da avaliação das candidatas a tradução (como correta, incorreta ou parcialmente correta). As traduções, como mencionadas anteriormente, são baseadas no número de candidatas consideradas como VP. Por exemplo, na primeira linha temos que dos 213 pares obtidos para o primeiro par de padrões, apenas 20 candidatas fonte foram consideradas VP. Dessas 20 VPs, 16 candidatas tiveram tradução corretamente recuperadas, 3 não foram tiveram traduções possíveis recuperadas e 1 teve uma tradução parcial recuperada.

Tabela 5.17: Precisão dos pares gerados pelo primeiro método de descoberta bilíngue

	# VP	# tradução correta	# tradução incorreta	# tradução parcial
1	90,9%	65,55%	20%	14,44%
2	79,38%	31,95%	29,89%	17,52%
3	9,38%	80%	15%	5%
4	100%	21,21%	61,61%	7,07%

Usando os valores da nossa Pontuação Final (ver seção 4.1) para avaliar os pares gerados pelo primeiro método, utilizamos o *ranking* final para calcularmos o MAP da lista de pares analisados em relação às candidatas fonte serem VP ou não e também em relação à candidata alvo ser considerada uma tradução, não ser considerada uma tradução ou ser considerada uma tradução parcial. Na tabela 5.18 verificamos esses valores. Quanto mais o valor está próximo de 1, significa que as candidatas mais bem pontuadas estavam sendo considerada como VP.

Tabela 5.18: MAP dos pares gerados pelo primeiro método de descoberta bilíngue

	# MAP VP	# MAP Traduções
1	0,9141	0,8707
2	0,7973	0,7815
3	0,1238	0,093
4	1	0,8696

Apesar do método usar a medida *Student T-Score* como componente do cálculo da Pontuação Final para cada par, decidiu-se não anotar se as candidatas alvo eram VP. O principal motivo

para essa decisão foi o fato de as construções verbo-suporte serem um desafio mesmo para quem tem a língua inglesa como língua nativa. Outro ponto importante é que chegamos a realizar experimentos com o padrão *Ficar + Adjetivo*, mas após consultas com especialistas descobrimos que verbos copulativos na verdade não fazem parte dos Verbos Construção Suporte e decidimos descartá-los da nossa análise.

Analisando o caso do padrão número 2 da tabela 5.17, verificamos que a precisão de traduções corretas foi de aproximadamente 21%, apesar da precisão de candidatas consideradas VP serem de 79%. Entretanto, se avaliarmos o valor de MAP, verificamos que esse valor alcança em torno de 79%, mostrando que mesmo que o método não tenha retornado tantas traduções boas, o cálculo da pontuação para cada par classificou traduções colocando as corretas em posições mais altas da lista e as incorretas em posições mais baixas.

Além disso, como verificamos que aplicar um padrão específico para *dar* poderia deixar muitas candidatas alvo de fora, pois a construção *dar* nem sempre é traduzida como *give*, "relaxamos" a busca dos padrões morfossintáticos do lado alvo. Entretanto, o pareamento de candidatas com padrões mais "relaxados" pode ter produzido muitas comparações desnecessárias, uma vez que as candidatas alvo poderiam aparecer na saída apenas por terem estrutura de verbo e substantivo. Ao olharmos a tabela 5.18, para a linha 4, verificamos que o resultado do MAP é de 86%. Como o cálculo do MAP leva em conta a pontuação final das candidatas (gerada pelo método), classificando as traduções e verificando quais delas foram consideradas como traduções fiéis pelos nossos juízes, verificamos que o cálculo da pontuação deixa as traduções menos prováveis em posições mais baixas na lista, fazendo com que o cálculo MAP final seja alto. Isto mostra como a estratégia de atribuir uma pontuação pode ajudar na hora de verificar as traduções, uma vez que o método dá uma pontuação baixa para pareamentos improváveis de candidatas.

Para o padrão das linhas 3 das tabelas mencionadas, o método mostrou um desempenho contrário. A precisão das traduções completas foi de 80%, entretanto o MAP foi de 9%. Neste caso, o método conseguiu ter um bom desempenho quanto ao método de traduções, entretanto, a pontuação final de cada pareamento não se mostrou efetivo para classificar as candidatas, fazendo com que muitas traduções corretas tivessem pontuação menor do que outras que não eram consideradas traduções.

5.4.2 Resultados dos experimentos com o segundo método de descoberta bilíngue

Para o segundo método de descoberta bilíngue, apenas a lista de candidatas em português foi gerada, uma vez que as candidatas à tradução são gerados pelos vetores bilíngues. Como o *Multivec* leva em consideração a frequência de ocorrência das palavras para gerar seus vetores, palavras que ocorriam pouco no córpus foram descartadas nessa geração. Assim, quando o segundo método de descoberta bilíngue tenta gerar as palavras similares para as palavras na candidatas fonte que são pouco frequentes, ele não encontra opções e essas candidatas precisam ser descartadas. Por isso, os valores da tabela 5.19 são diferentes (e menores do que os) da tabela 5.14.

Tabela 5.19: Número de candidatas retornadas após a aplicação dos padrões morfossintáticos no córpus em português desconsiderando aquelas para as quais não foram gerados vetores de palavras

	Padrões em português	# candidatas diferentes	# total de candidatas
1	Realizar + SUBSTANTIVO	190	567
2	[Fazer | Tomar] + SUBSTANTIVO	782	2948
3	[Entrar| Colocar] + SUBSTANTIVO	472	1141
4	Dar + SUBSTANTIVO	358	1454

A tabela 5.20 mostra os números de pares retornados na aplicação do segundo método. Na mesma tabela, na segunda coluna, mostramos o número de pares analisados pelos nossos juízes. Para a escolha das candidatas a serem analisadas nesse método, tentamos usar os pares que continham as mesmas candidatas fonte do método anterior. Entretanto, os números mudam pelo fato de que, por exemplo, a candidata *realizar processo* pode ter aparecido 5 vezes em combinações diferentes no método 1, mas no segundo apareceu apenas 2. Isso influencia no número total de pares analisados nos dois experimentos.

Tabela 5.20: Número de pares gerados pelo segundo método de descoberta bilíngue

	Padrões em português	# pares	# pares avaliados
1	Realizar + SUBSTANTIVO	318	130
2	[Fazer | Tomar] + SUBSTANTIVO	1583	139
3	[Entrar| Colocar] + SUBSTANTIVO	754	164
4	Dar + SUBSTANTIVO	689	45

A tabela 5.21 mostra os resultados que obtivemos após a avaliação das candidatas. A primeira coluna mostra o número de candidatas analisadas que foram consideradas EMs em relação ao número total de candidatas anotadas. Nas colunas restantes temos a avaliação da candidata alvo como tradução da candidata fonte, baseado no número de candidatas fontes consideradas como EM.

Tabela 5.21: Avaliação dos pares gerados pelo segundo método de descoberta bilíngue

	# VP	# tradução correta	# tradução incorreta	# tradução parcial	# sem tradução
1	88,46%	71,30%	6,95%	20%	1,73%
2	75,53%	33,04%	11,42%	40%	12,38%
3	12,80%	38,09%	9,52%	52,38%	0%
4	100%	44,44%	11,11%	28,88%	15,55%

O número de candidatas fonte sendo consideradas candidatas a EM na tabela 5.21 aqui não é o mesmo que no experimento anterior, com o primeiro método, pelo fato de que algumas candidatas de um mesmo padrão no primeiro método podem não ter sido alinhadas com outra candidata alvo, ficando de fora do nosso cálculo final. Algumas candidatas fonte são consideradas como EM, muitas vezes dependendo do seu contexto, por isso nossos juízes tinham as sentenças originais do córpus como apoio. Por exemplo, a expressão *entrar [em] circulação* pode ocorrer em contextos onde ela é uma EM, como em *a moeda entrou em circulação*, ou em contextos onde ela não é uma EM, como em *o medicamento entrou na circulação sanguínea*. Apenas no primeiro caso a expressão será considerada EM pois a segunda infere um significado literal da ação.

Para o segundo método, foi necessário fazer o cálculo do número de vezes em que uma candidata fonte não possuía nenhuma candidata alvo (tradução possível na sentença paralela). Como já mencionado anteriormente, o método que usa *embeddings* para gerar traduções possui suas limitações pelo fato de que as traduções são mapeadas de acordo com a frequência que elas ocorrem. Felizmente, foram poucos os casos nos quais o método não conseguiu inferir uma tradução.Analisando o resultado da precisão de traduções dos 4 padrões, a porcentagem de vezes que o método não inferiu nenhuma tradução foram de 0%, 1%, 12% e 15%. Uma possível extensão deste segundo método seria propor uma alternativa para esses casos onde as *embeddings* não conseguem auxiliar na recuperação de possíveis traduções.

Em relação às traduções recuperadas, temos uma alta taxa de acerto para traduções corretas para o padrão 1 (71%), mas o mesmo não ocorre para os demais padrões: sendo de apenas 36% para o padrão 2, 38% para o padrão 3 e de 44% para o padrão 4. Contudo, quando consideramos também as traduções parciais essas taxas sobem para 91%, 76%, 90% e 73% para os padrões 1, 2, 3 e 4, respectivamente. Isso significa que ao trabalharmos em mais formas de refinar nosso experimento poderíamos tornar casos de tradução parcial em traduções totalmente corretas.

Não calculamos o MAP para essa segunda abordagem pois este metódo não gera pontuação final para cada pareamento fazendo com que seja inviável ordenar os pares.

5.5 Discussões

Após a análise dos resultados e de todo o trabalho desenvolvido, separamos alguns pontos importantes e que serão discutidos nesta seção a fim de que possamos verificar as vantangens e desvantagens de cada método, com o intuito também de levantarmos questões pertinentes à eficiência dos métodos e como eles podem ser melhorados. Vamos começar, então, enumerando as questões levantadas durante o desenvolvimento do método de descoberta de EMs como um todo:

1. O método gera alguma pontuação final para que possamos criar algum *ranking* de candidatas a EM ou, então, essa pontuação pode ser levada em conta para considerarmos uma candidata como EM sem a necessidade de uma análise feita por um perito?

   Podemos dizer que somente o primeiro método lida com essa questão. O cálculo final do primeiro método tem o objetivo de dar ao usuário uma pontuação que ajude a decidir o limiar de avaliação dentro de um *ranking* de candidatas com suas respectivas traduções, por exemplo, se numa lista de 100 candidatas, o usuário avaliaria as 50 primeiras com melhor pontuação. O fato de as candidatas alvo pareadas serem buscadas de forma monolíngue também faz com que tanto a candidata fonte quanto a candidata alvo possuam pontuações anteriores que mostram as chances de elas, de forma separada, serem consideradas EMs. O segundo método não possui nenhum cálculo que dê uma pista da candidata e sua tradução pois apenas gera uma tradução baseada no método de *embeddings* usando alguns filtros já mencionados. No caso do segundo método, podemos usar a probabilidade de um par ter acontecido no córpus originalmente como um parâmetro de decisão. Por exemplo, se o nosso método recuperou *dar exemplo* como *give example* 8 vezes, em diferentes frases, isso pode indicar que nosso método está recuperando essa candidata a tradução de forma correta.

2. O método leva em conta o fato de que queremos gerar traduções em uma língua em que não dominamos?

   Levantamos o ponto de que podem acontecer casos em que queremos inferir traduções de EMs mas não temos muito domínio da língua alvo. O primeiro método infelizmente exige que façamos uma busca de padrões do lado fonte e alvo. Nesse caso, se não dominamos a língua alvo, a ideia é de que coloquemos padrões "flexíveis", como por exemplo procurar todos os verbos seguidos de substantivos. Já o segundo método, como ele infere uma tradução usando o método de *embeddings*, o usuário não precisa conhecer a língua alvo.

3. O método lida com o fato de que algumas candidatas a EM fonte são traduzidas como apenas uma palavra na língua alvo? Ele considera que a tradução de uma expressão fonte nem sempre é a tradução literal da mesma? Esse item levanta dois pontos ao mesmo tempo. Para o primeiro ponto justificamos que, ao usarmos padrões fixos para o primeiro método, deixamos de fora situações que não se encaixam nesse padrão. Se fizermos uma busca por apenas verbos numa língua alvo, teremos uma sobrecarga de processamento, pois a extração monolíngue alvo retornaria todos os verbos de um córpus inteiro. Já o segundo método consegue lidar com essa questão, usando as etiquetas morfossintáticas como pista se estamos lidando com um verbo, por exemplo. Para o ponto de uma candidata nem sempre possuir traduções literais, o primeiro método só conseguiria talvez lidar com isso se o padrão monolíngue alvo fosse flexível, onde usaríamos o cálculo de similaridade para verificar se as duas candidatas têm chances de serem tradução uma da outra. Como o segundo método não espera um padrão monolíngue alvo, apenas o método das *embeddings* irá mostrar o que já está mapeado de acordo com o córpus fonte. Entretanto, para os dois métodos, o método de vetor de palavras nem sempre consegue retornar traduções ou calcular a similaridade de forma satisfatória, uma vez que certas palavras podem ser traduzidas originalmente no córpus paralelo com baixa frequência.

4. O método gera muito ruído? O método gera pares que poderiam ter sido evitados na saída por algum filtro? O primeiro método pode gerar muitos pareamentos desnecessários se o padrão de busca monolíngue for muito flexível, como aconteceu em alguns de nossos experimentos. Uma alternativa para esse problema é colocar um limiar para escolher as candidatas a serem avaliadas. Já o segundo método não retorna muito ruído, pois ele apenas realiza a tradução das candidatas fontes da lista de entrada, entretanto ele pode não conseguir inferir uma tradução.

   Para os dois métodos usamos ferramentas automáticas para a etiquetação morfossintática a e geração de vetores bilíngues de palavras, as quais são passíveis de erros e, por isso, nem sempre acertam todas as classes de palavras nem mapeiam as traduções de palavras com baixa frequência no córpus. O segundo método não tem uma pontuação da chance das traduções serem consideradas EMs, o que faz com que não tenhamos mais pistas se essa tradução poderia ser considerada uma tradução de EM. O segundo método também deve ser melhorado quanto aos filtros usados.

5. Qual seria o trabalho necessário para adaptar os métodos para outras categorias de EM? Até o momento, usamos indícios linguísticos para organizarmos as traduções encontradas usando as premissas de que traduções de construção verbo suporte são traduzidas

como *verbo + substantivo* ou apenas verbo. O método está apenas lidando com EMs desse tipo, não estamos lidando com traduções de colocações, expressões idiomáticas etc. Para os método ser capaz de lidar com outras categorias seria necessário um estudo de como as traduções das mesmas se comportam ao longo de um córpus e aplicando essas informações no método.

Ambos os métodos desenvolvidos, portanto, são capazes de realizar a descoberta bilíngue de EMs. A pontuação final gerada pelo primeiro método possibilita um ordenamento do pareamento de candidatas gerando um limiar para avaliação manual das mesmas. Entretanto, o segundo método apresentou menos limitações que o primeiro, pelo fato de que vetores bilíngues são capazes de recuperar traduções oficiais de candidatas a partir do córpus paralelo, de forma que consiga abranger os casos nos quais uma candidata fonte foi traduzida apenas como um verbo, por exemplo. Para realizar a extração monolíngue a fim de gerar as listas de entrada para o primeiro método, é necessário um conhecimento prévio da língua alvo, o que não ocorre no segundo método, pois ele apenas necessita como entrada uma lista de candidatas fonte. Em relação à taxa de precisão, o segundo método possui índices altos para acertos parciais, mostrando que é possível melhorar a precisão de tradução completa se seus filtros forem melhor trabalhados. Embora os dois métodos apresentem vantagens e desvantagens, o segundo método é o escolhido para ser disponibilizado pois é capaz de lidar com as particularidades de traduções de EMs, como o fato de nem sempre serem traduzidas como uma expressão e nem traduzidas de forma literal.

Capítulo 6

CONSIDERAÇÕES FINAIS

Expressões Multipalavras são um desafio atual para a área de PLN e existem diferentes métodos automáticos propostos para descobri-las e tratá-las nos textos. Por não possuírem regras específicas para sua formação, é necessário o uso de métodos estatísticos e outras fontes de informação, como córpus bilíngues, para sua descoberta e processamento.

Vimos, em trabalhos desde Zarrieß e Kuhn (2009) a trabalhos como o de Garcia, García-Salido e Alonso-Ramos (2017), que a descoberta bilíngue trouxe muitos ganhos para a descoberta de EMs por fazer uma comparação entre mais idiomas, tentando extrair informações que possam ser úteis para melhores desempenhos da descoberta. Entretanto, em muitos experimentos notamos que a tradução literal de uma candidata de um lado fonte para um lado alvo pode ser problemática. Quando estamos falando de colocações ou expressões idiomáticas, muitas vezes as correspondências nas línguas não constituem traduções literais ou o número de palavras em ambos os lados não é igual. Além disso, há casos em que temos construções verbo suporte (como *dar ênfase*) para as quais a tradução no outro idioma é apenas uma palavra (em inglês, *emphasize*). Por essa razão, o uso de traduções estatísticas e dicionários e até mesmo o alinhamento de palavras pode prejudicar a descoberta. Isso acontece pois, como mencionado, traduções com o uso de dicionário tendem a inferir traduções do tipo literal, e de acordo com o que foi descoberto até o momento, nem sempre EM são traduzidas de forma literal, principalmente expressões idiomáticas que possuem no nome a força que o idioma traz para suas construções. A ideia preliminar então do nosso método era desenvolver técnicas de descoberta bilíngue em textos paralelos sem depender de recursos léxicos externos e verificar se era possível usarmos apenas o alinhamento de sentenças para essa descoberta.

O objetivo deste trabalho foi verificar a possibilidade de gerar automaticamente traduções de expressões multipalavras tendo como objeto de estudo as construções verbo-suporte. Para tanto, a hipótese de pesquisa era a de que é possível recuperar o paralelismo entre EMs e suas

traduções usando vetores de palavras bilíngues. Em outras palavras, a estratégia adotada neste trabalho se baseou no alinhamento em nível de sentença ao invés de alinhamento de palavras.

Durante o desenvolvimento deste trabalho, verificamos que os vetores bilíngues eram úteis não apenas como uma forma de compararmos palavras de um lado fonte e o lado alvo para inferirmos o nível de similaridade, como descobrimos também, que os vetores são úteis também para gerar possíveis traduções. Por exemplo, ao invés de apenas verificarmos o quanto as palavras *realizar* e *carry out* eram similares (estratégia usada na abordagem da primeira proposta), os vetores bilíngues também foram capazes de retornar palavras como *carry out* ou *conduct* como de *realizar* (estratégia usada na segunda proposta). A grande vantagem dos vetores bilíngues, portanto, é o fato de que eles podem mapear as palavras de acordo com a ocorrência nos textos bilíngues, ou seja, podem inferir traduções de acordo com o uso das palavras em contextos nos textos. Dicionários bilíngues, por sua vez, possuem traduções fixas, diferentemente dos vetores que conseguem mapear palavras de acordo com o uso. É importante lembrar que as línguas estão sempre mudando e palavras novas vão aparecendo no nosso vocabulário. Muitos dicionários têm suas traduções fixas e talvez já datadas de algum tempo, porém usando textos bilíngues atualizados para mapear as palavras e inferirmos traduções, podemos ter um processo mais dinâmico. Por exemplo, cada vez que trocamos os textos bilíngues, podemos ter tipos diferentes de tradução e de acordo com o uso de termos. Se estamos mais interessados em inferir traduções e analisar EMs de textos jornalísticos, ou se estamos interessados em textos de música, os vetores bilíngues são ideais para a descoberta bilíngue. Entretanto, verificamos que os vetores bilíngues tem um grande alcance mas ainda não resolvem problemas como traduções pontuais e menos frequentes. Por exemplo, *realizar sonho*, onde foi traduzido no texto paralelo uma vez como *fulfill dream*. No caso, o verbo *realizar* ser traduzido como *fulfill*, de maneira pontual no texto, fazendo com que os vetores bilíngues não conseguissem sugerir *fulfill* como palavra alvo similar para *realizar*.

Durante o desenvolvimento do segundo método de descoberta bilíngue, verificamos que classes gramaticais acabaram por atuar como indícios linguísticos na hora de realizarmos as traduções das candidatas. Fazendo uso das classes gramaticais, conseguiríamos mapear os casos em que uma candidata fonte era traduzida apenas como verbo, e se uma tradução não fosse do tipo *verbo + substantivo*, aplicávamos um filtro para obtermos essa estrutura. Contudo, esses resultados foram verificados, até o momento, apenas para construções verbo-suporte e não podemos afirmar se características linguísticas ajudariam na tradução nos demais tipos de EM. Como trabalho futuro, acreditamos que deva-se explorar se essas premissas de uso de vetores bilíngues e uso de classes gramaticais podem ser aplicadas para descoberta bilíngue que envolva outros tipos de EMs.

A descoberta de EMs portanto mostra-se uma tarefa difícil por estar diretamente ligada com o uso da língua e nem sempre atributos linguísticos são úteis na descoberta. Como os vetores bilíngues conseguem mapear e inferir traduções a partir de textos paralelos, esse método está mais diretamente ligado ao uso da língua pelas pessoas que escrevem tais textos, do que o uso de dicionários bilíngues, em que certas traduções já foram atribuídas e fixadas em um léxico. Acreditamos que os dois métodos desenvolvidos neste trabalho usufruem dos recursos de vetores bilíngues. Apesar de apenas o primeiro método de descoberta bilíngue desenvolvido contar com uma pontuação final que nos ajudaria a criar uma classificação a partir de um *ranking*, o segundo método é capaz de retornar casos onde uma EM não foi traduzida na mesma estrutura e sendo capaz de inferir traduções diretamente dos vetores bilíngues, além de não ser necessário a descoberta monolíngue alvo, uma vez que o método não depende de conhecimento prévio de uma língua alvo. O segundo método portanto é o que possui mais indícios para ser usado como método de descoberta bilíngue para Construções Verbo Suporte até o momento e será disponibilizado na plataforma *Github* do LALIC[1].

6.1 Trabalhos Futuros

Apesar de o segundo método ter se mostrado mais completo, uma combinação dos métodos poderia ser investigada como trabalho futuro. Mais especificamente, poderia-se investigar maneiras de unir elementos do primeiro método (como a pontuação final) com elementos do segundo método, que gera traduções de apenas uma palavra.

Outra possível extensão deste trabalho diz respeito ao tratamento de outros tipos de EMs. Para tanto, será necessária uma investigação de como o método poderia ser adaptado a outros casos. Por exemplo, seria interessante investigar como outros tipos de EMs são traduzidas para outra língua com estratégias propostas para cada tipo em especial, se for o caso.

Em relação aos recursos utilizados pelo BiDiMWEToolkit, poderia-se também testar o uso combinado de léxicos externos a fim de validar casos onde as informações linguísticas não são suficientes para uma descoberta bilíngue eficiente baseada apenas nos vetores de palavras.

[1] https://github.com/LALIC-UFSCar

REFERÊNCIAS

ACOSTA, O. C.; VILLAVICENCIO, A.; MOREIRA, V. P. Identification and treatment of multiword expressions applied to information retrieval. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. [S.l.], 2011. p. 101–109.

ARMENTANO-OLLER, C. et al. Open-source portuguese-spanish machine translation. In: SPRINGER. *PROPOR*. [S.l.], 2006. p. 50–59.

ATTIA, M. et al. Automatic extraction of arabic multiword expressions. In: *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*. [S.l.: s.n.], 2010. p. 19–27.

AZIZ, W.; SPECIA, L. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In: *STIL 2011*. Cuiabá, MT: [s.n.], 2011.

BAI, M.-H. et al. Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. [S.l.], 2009. p. 478–486.

BALDWIN, T.; KIM, S. N. Multiword expressions. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). *Handbook of Natural Language Processing*. 2nd. ed. Boca Raton: CRC Press, 2010.

BENGIO, Y. et al. A neural probabilistic language model. *Journal of machine learning research*, v. 3, n. Feb, p. 1137–1155, 2003.

BÉRARD, A. et al. Multivec: a multilingual and multilevel representation learning toolkit for nlp. In: *The 10th edition of the Language Resources and Evaluation Conference (LREC)*. [S.l.: s.n.], 2016.

BESANÇON, R. et al. Lima: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In: *LREC*. [S.l.: s.n.], 2010.

BIBER, D. et al. *Longman grammar of spoken and written English*. [S.l.]: MIT Press Cambridge, MA, 1999.

BOUAMOR, D.; SEMMAR, N.; ZWEIGENBAUM, P. Identifying bilingual multi-word expressions for statistical machine translation. In: *LREC*. [S.l.: s.n.], 2012. p. 674–679.

BRANTS, T.; FRANZ, A. Web 1t 5-gram version 1. Linguistic Data Consortium, Philadelphia, 2006.

BREEN, J. Building an electronic japanese-english dictionary. In: *Japanese Studies Association of Australia Conference*. [S.l.: s.n.], 1995.

BROWN, P. F. et al. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, MIT Press, v. 19, n. 2, p. 263–311, 1993.

CALZOLARI, N. et al. Towards best practice for multiword expressions in computational lexicons. In: *LREC*. [S.l.: s.n.], 2002.

CARPUAT, M.; DIAB, M. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, 2010. p. 242–245. Disponível em: <http://www.aclweb.org/anthology/N10-1029>.

CASELI, H. M. de et al. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, Springer, v. 44, n. 1-2, p. 59–77, 2010.

CHOUEKA, Y. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In: *RIAO 88:(Recherche d'Information Assistée par Ordinateur). Conference*. [S.l.: s.n.], 1988. p. 609–623.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. *Computational linguistics*, MIT Press, v. 16, n. 1, p. 22–29, 1990.

CONSTANT, M. et al. Multiword expression processing: A survey. *Computational Linguistics*, MIT Press, v. 43, n. 4, p. 837–892, 2017.

DENKOWSKI, M.; LAVIE, A. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the sixth workshop on statistical machine translation*. [S.l.], 2011. p. 85–91.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, American Psychological Association, v. 76, n. 5, p. 378, 1971.

FORCADA, M. L. et al. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, Springer, v. 25, n. 2, p. 127–144, 2011.

GARCIA, M.; GAMALLO, P. Yet another suite of multilingual nlp tools. In: SPRINGER. *International Symposium on Languages, Applications and Technologies*. [S.l.], 2015. p. 65–75.

GARCIA, M.; GARCÍA-SALIDO, M.; ALONSO-RAMOS, M. Using bilingual word-embeddings for multilingual collocation extraction. In: *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. [S.l.: s.n.], 2017. p. 21–30.

HOANG, H. H.; KIM, S. N.; KAN, M.-Y. A re-examination of lexical association measures. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. [S.l.], 2009. p. 31–39.

HOFLAND, K. A program for aligning english and norwegian sentences. *Research in humanities computing*, v. 5, p. 165–178, 1996.

ITAI, A.; WINTNER, S. Language resources for hebrew. *Language Resources and Evaluation*, Springer, v. 42, n. 1, p. 75–98, 2008.

JACKENDOFF, R. Twistin'the night away. *Language*, JSTOR, p. 534–559, 1997.

KOEHN, P. et al. Moses: Open source toolkit for statistical machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. [S.l.], 2007. p. 177–180.

KORKONTZELOS, I.; MANANDHAR, S. Detecting compositionality in multi-word expressions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. [S.l.], 2009. p. 65–68.

KRENN, B.; EVERT, S. et al. Can we do better than frequency? a case study on extracting pp-verb collocations. In: *Proceedings of the ACL Workshop on Collocations*. [S.l.: s.n.], 2001. p. 39–46.

KUDO, T.; MATSUMOTO, Y. Japanese dependency analysis using cascaded chunking. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *proceedings of the 6th conference on Natural language learning-Volume 20*. [S.l.], 2002. p. 1–7.

LISON, P.; TIEDEMANN, J. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: *LREC*. [S.l.: s.n.], 2016.

LUONG, T.; PHAM, H.; MANNING, C. D. Bilingual word representations with monolingual quality in mind. In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. [S.l.: s.n.], 2015. p. 151–159.

MEL'ČUK, I. Collocations and lexical functions. *Phraseology. Theory, analysis, and applications*, Clarendon Press, Oxford, p. 23–53, 1998.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

NIVRE, J.; HALL, J.; NILSSON, J. Maltparser: A data-driven parser-generator for dependency parsing. In: *Proceedings of LREC*. [S.l.: s.n.], 2006. v. 6, p. 2216–2219.

OCH, F. J.; NEY, H. A systematic comparison of various statistical alignment models. *Computational Linguistics*, v. 29, n. 1, p. 19–51, 2003.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 40th annual meeting on association for computational linguistics*. [S.l.], 2002. p. 311–318.

PEARSON, K. *On the theory of contingency and its relation to association and normal correlation; On the general theory of skew correlation and non-linear regression*. [S.l.]: Cambridge University Press, 1904.

PECINA, P. Lexical association measures and collocation extraction. *Language resources and evaluation*, Springer, v. 44, n. 1-2, p. 137–158, 2010.

PENNINGTON, J.; SOCHER, R.; MANNING, C. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543.

PEREIRA, L. et al. Identifying collocations using cross-lingual association measures. *EACL 2014*, p. 109, 2014.

RAMISCH, C. A generic framework for multiword expressions treatment: from acquisition to applications. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of ACL 2012 Student Research Workshop*. [S.l.], 2012. p. 61–66.

RAMISCH, C. *Multiword Expressions Acquisition: A Generic and Open Framework*. [S.l.]: Springer, 2015.

RAMISCH, C.; VILLAVICENCIO, A.; BOITET, C. mwetoolkit: a Framework for Multiword Expression Identification. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valetta, Malta: European Language Resources Association, 2010. ISBN 2-9517408-6-7.

RAMISCH, C.; VILLAVICENCIO, A.; KORDONI, V. Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing (TSLP)*, ACM, v. 10, n. 2, p. 3, 2013.

REN, Z. et al. Improving statistical machine translation using domain bilingual multiword expressions. In: *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. Singapore: Association for Computational Linguistics, 2009. p. 47–54. Disponível em: <http://www.aclweb.org/anthology/W/W09/W09-2907>.

SAG, I. A. et al. Multiword expressions: A pain in the neck for nlp. In: SPRINGER. *International Conference on Intelligent Text Processing and Computational Linguistics*. [S.l.], 2002. p. 1–15.

SALEHI, B.; COOK, P. Predicting the compositionality of multiword expressions using translations in multiple languages. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. [S.l.: s.n.], 2013. v. 1, p. 266–275.

SALEHI, B.; COOK, P.; BALDWIN, T. A word embedding approach to predicting the compositionality of multiword expressions. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2015. p. 977–983.

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

SAVARY, A. et al. The parseme shared task on automatic identification of verbal multiword expressions. In: *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. [S.l.: s.n.], 2017. p. 31–47.

SCHMID, H. Treetagger. *TC project at the Institute for Computational Linguistics of the University of Stuttgart*, 1994.

SMITH, A. Breaking bad: Extraction of verb-particle constructions from a parallel subtitles corpus. In: *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. [S.l.: s.n.], 2014. p. 1–9.

SOCHER, R. et al. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2013. v. 1631, p. 1642.

TIEDEMANN, J. Parallel data, tools and interfaces in opus. In: *LREC*. [S.l.: s.n.], 2012. v. 2012, p. 2214–2218.

TSVETKOV, Y.; WINTNER, S. Automatic acquisition of parallel corpora from websites with dynamic content. In: *LREC*. [S.l.: s.n.], 2010.

TSVETKOV, Y.; WINTNER, S. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, Cambridge University Press, v. 18, n. 4, p. 549–573, 2012.

TSVETKOV, Y.; WINTNER, S. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics*, MIT Press, v. 40, n. 2, p. 449–468, 2014.

WILKS, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, JSTOR, v. 9, n. 1, p. 60–62, 1938.

ZAMPIERI, N. et al. Veyn at parseme shared task 2018: Recurrent neural networks for vmwe identification. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. [S.l.: s.n.], 2018. p. 290–296.

ZARRIESS, S.; KUHN, J. Exploiting translational correspondences for pattern-independent mwe identification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. [S.l.], 2009. p. 23–30.

ZOU, W. Y. et al. Bilingual word embeddings for phrase-based machine translation. In: *EMNLP*. [S.l.: s.n.], 2013. p. 1393–1398.

GLOSSÁRIO

EM – *Expressão Multipalavras*

PLN – *Processamento de Linguagem Natural*

VP – *Verdadeiro Positivo*