UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

KASSIUS VARGAS PRESTES

Extração Multilíngue de Termos Multipalavra em Corpora Comparáveis

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Profa. Dra. Aline Villavicencio

Porto Alegre
2015

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Prestes, Kassius Vargas

Extração Multilíngue de Termos Multipalavra em Corpora Comparáveis / Kassius Vargas Prestes. – Porto Alegre: PPGC da UFRGS, 2015.

84 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Ciência da Computação, Porto Alegre, BR–RS, 2015. Orientador: Aline Villavicencio.

1. Processamento de linguagem natural. 2. Extração de termos. 3. Alinhamento multilíngue. 4. Alinhamento corpora comparável. 5. Corpus. I. Villavicencio, Aline. II. Título.

AGRADECIMENTOS

Agradeço aos meus pais e à minha irmã por sempre me apoiarem em todas as etapas da minha vida.

Agradeço aos projetos que apoiaram e apoiam o grupo de Processamento de Linguagem Natural na UFRGS, principalmente o projeto Pronem e à professora Renata Vieira.

Agradeço à Bianca, à Érica e à Liz que me ajudaram na avaliação dos resultados do meu trabalho, sem os quais ele não poderia ser concluído.

Agradeço ao Rodrigo Wilkens que me acompanha desde a graduação me ajudando e orientando.

Agradeço à minha orientadora Aline Villavicencio que também me acompanha desde a graduação por todas as dicas, idéias e apoio durante todo esse período.

Também agradeço ao apoio dos projetos Cameleon (Capes-Cofecub 707/11) e Fapergs-CNRS-INRIA AIM-West.

“It ain’t about how hard you hit. It’s about how hard you can get hit and keep moving forward.”

Rocky Balboa (STALLONE, 2006)

“It’s not who you are underneath, it’s what you do that defines you.”

Batman Begins (NOLAN, 2005)

Multilingual Extraction of Multiword Terms in Comparable Corpora

ABSTRACT

This work investigates techniques for multiword term extraction from comparable corpora, which are sets of texts in two (or more) languages about the same topic. Term extraction, specially multiword terms is very important to help the creation of terminologies, ontologies and the improvement of machine translation. In this work we use a comparable corpora Portuguese/English and want to find terms and their equivalents in both languages. To do this we start with separate term extraction for each language. Using morphossintatic patterns to identify n-grams (sequences of n words) most likely to be important terms of the domain. From the terms of each language, we use their context, i. e., the words that occurr around the term to compare the terms of different languages and to find the bilingual equivalents. We had as main goals in this work identificate monolingual terms, apply alignment techniques for Portuguese and evaluate the different parameters of size and type (used PoS) of window to the context extraction. This is the first work to apply this methodology to Portuguese and in spite of the lack of lexical and computational resources (like bilingual dictionaries and parsers) for this language, we achieved results comparable to state of the art in French/English.

Keywords: Natural language processing, term extraction, multilingual alignment, comparable corpora alignment, corpus.

RESUMO

Este trabalho investiga técnicas de extração de termos multipalavra a partir de corpora comparáveis, que são conjuntos de textos em duas (ou mais) línguas sobre o mesmo domínio. A extração de termos, especialmente termos multipalavra é muito importante para auxiliar a criação de terminologias, ontologias e o aperfeiçoamento de tradutores automáticos. Neste trabalho utilizamos um corpus comparável português/inglês e queremos encontrar termos e seus equivalentes em ambas as línguas. Para isso começamos com a extração dos termos separadamente em cada língua, utilizando padrões morfossintáticos para identificar os n-gramas (sequências de n palavras) mais prováveis de serem termos importantes para o domínio. A partir dos termos de cada língua, utilizamos o contexto, isto é, as palavras que ocorrem no entorno dos termos para comparar os termos das diferentes línguas e encontrar os equivalentes bilíngues. Tínhamos como objetivos principais neste trabalho fazer a identificação monolíngue de termos, aplicar as técnicas de alinhamento para o português e avaliar os diferentes parâmetros de tamanho e tipo (PoS utilizados) de janela para a extração de contexto. Esse é o primeiro trabalho a aplicar essa metodologia para o Português e apesar da falta de alguns recursos léxicos e computacionais (como dicionários bilíngues e parsers) para essa língua, conseguimos alcançar resultados comparáveis com o estado da arte para trabalhos em Francês/Inglês.

Palavras-chave: Processamento de linguagem natural. extração de termos. alinhamento multilíngue. alinhamento corpora comparável. corpus.

LISTA DE ABREVIATURAS E SIGLAS

EM	Expressão Multipalavra
IDF	Inverse Document Frequency
LL	Log Likelihood
LLR	Log Likelihood Ratio
MAP	Mean Average Precision
MI	Mutual Information
MWE	Multiword Expression
MWT	Multiword Term
NLP	Natural Language Processing
NP	Noun Phrase
PLN	Processamento de Linguagem Natural
PMI	Pointwise Mutual Information
PoS	Part-of-Speech
PS	Poisson Stirling
TMI	True Mutual Information
UAP	Uninterpolated Average Precision
WEKA	Waikato Environment for Knowledge Analysis

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Atualmente a informação na web cresce em uma velocidade muito grande, porém grande parte dessa informação está acessível apenas em poucas línguas, principalmente em inglês. Para esta informação vir a ser utilizada de maneira adequada em sistemas computacionais, tais como, agentes conversacionais, sistemas de perguntas e respostas e de tradução automática são necessários recursos lexicais e ontológicos que formalizem e unifiquem o conhecimento contido na web. Existem recursos disponíveis para domínios específicos como medicina, mercado imobiliário e petróleo, porém são necessários para os mais diversos domínios e em diversas línguas. Uma alternativa à construção manual desses recursos, que é um processo caro e que demanda tempo e conhecimento especializado, é o desenvolvimento de métodos automáticos de identificação de termos.

Uma das unidades básicas, que não são muito bem tratadas nos sistemas computacionais, mas são necessárias para a criação desse tipo de recurso são as **expressões multipalavra** (EM, ou em inglês *multiword expressions*, ou MWEs) que são um conceito genérico que descrevem um grande número de fenômenos linguísticos distintos, mas relacionados, como:

- compostos nominais (*international symposium*, desenvolvimento sustentável),
- expressões idiomáticas (*kick the bucket*, sem pé nem cabeça),
- *phrasal verbs* (*take off, break in*),
- termos compostos (*software engineering*, modelo conceitual) e outros.

Definições alternativas de **expressão multipalavra** são muito bem discutidas por Ramisch (2012), que explica as características de cada uma das nomenclaturas que este fenômeno linguístico já teve ao longo da história e as características particulares de cada uma delas, como "colocações"(*collocations*), expressões idiomáticas (*idiom*) e expressão multipalavra (*multiword expression*). Aqui adotaremos uma definição genérica de Calzolari (2002):

Definição 1.1. Expressão Multipalavra *É uma sequência de palavras que atua como uma unidade em algum nível de análise linguística.*

Uma definição relacionada e muito importante para este trabalho, no contexto de um domínio específico é a de **termo**. De acordo com Krieger e Finatto (2004), para um especialista, termos são a representação do conhecimento em uma área específica, ou seja, são unidades léxicas que representam conceitos abstratos de um domínio.

A seguir apresentamos duas definições de termos.

Definição 1.2. Termo *Um termo é uma unidade léxica que possui um significado não ambíguo quando usado em um texto de domínio específico. Termos são frequentemente tratados como uma unidade, mesmo que sejam compostos por mais de uma palavra ortográfica. Um conjunto de termos específicos forma um léxico especializado ou uma terminologia.*(KRIEGER; FINATTO, 2004) *Uma terminologia é um léxico especializado correspondente a um conjunto*

de palavras que caracterizam uma linguagem especializada de um domínio.(CABRÉ; SAGER, 1999)

Existem algumas diferenças entre MWEs e termos, enquanto MWEs são compostas sempre por mais de uma palavra, termos podem ser palavras simples. MWEs ocorrem em textos de linguagem técnica/científica e também de domínio geral, enquanto termos ocorrem somente no primeiro tipo. A intersecção entre esses dois conceitos, os **termos multipalavra**, que são termos compostos de duas ou mais palavras ortográficas.

Definição 1.3. Termo Multipalavra *É uma unidade lexical composta de duas ou mais palavras cujo significado não pode ser inferido por um não-especialista a partir de suas partes porque ele depende de uma área específica e do conceito que ele descreve.* (SANJUAN et al., 2005) (FRANTZI; ANANIADOU; TSUJII, 1998)

Neste trabalho focaremos nos termos multipalavra pois estes têm uma semântica mais sólida e distintiva do que termos de somente uma palavra (DRYMONAS, 2009). Outro fator que contribui com a importância da utilização de termos multipalavra para a construção de terminologias é que esses termos multipalavra constituem de aproximadamente 50% do vocabulário de uma pessoa (JACKENDOFF, 1997) e, em um vocabulário de domínio específico esse percentual chega a 70% (KRIEGER; FINATTO, 2004), com a maioria dos termos sendo substantivos e compostos nominais.

Estudaremos duas tarefas relacionadas aos termos multipalavra, a primeira é a Extração Automática de termos a partir de corpora, e a segunda é o Alinhamento dos termos correspondentes em duas línguas. A Extração de Termos, também conhecida como Reconhecimento Automático de Termos ou Mineração de Termos, é uma tarefa de muitas aplicações de Linguagem Natural, como construção automática de dicionários e aprendizado de ontologias. Esse processo envolve a extração e a filtragem dos candidatos a termos, com o objetivo principal de determinar quando uma palavra ou uma sequência de palavras, é ou não um termo característico de um domínio.

A questão principal na Extração pode ser dividida em duas noções críticas nessa área: *unithood* e *termhood*. Definidas formalmente por Kageura e Umino (1996), *unithood* é o "grau de força ou estabilidade das combinações sintagmáticas e locuções"[1] e *termhood* é o "grau que uma unidade linguística é relacionada a um conceito de domínio específico". *Unithood* só é relevante para termos complexos, termos multipalavra, enquanto *termhood* serve para termos simples e complexos.

A segunda tarefa é o Alinhamento de Termos, no qual, a partir de uma lista de Termos do mesmo domínio, em duas línguas, queremos encontrar as equivalências entre eles. As tarefas de Extração e Alinhamento, juntas, podem ser chamadas de Extração Multilíngue, em que temos como entradas dois (ou mais) conjuntos de textos, um em cada língua e temos como resultado uma lista de termos e seus equivalentes em cada língua. No contexto multilíngue, queremos

[1]Do original, *"degree of strenght or stability of syntagmatic combinations and collocations"*

extrair termos, especialmente compostos, chamados de termos multipalavra, para criarmos e enriquecermos diversos recursos léxicos como dicionários multilíngues e ontologias, que por sua vez podem ajudar a melhorar a performance de tradutores automáticos que são muito usados hoje em dia.

O processo de Extração Multilíngue é feito a partir de um corpus[2]. Existem dois tipos de corpus usados nesta tarefa, os paralelos e os comparáveis. Um corpus paralelo contém textos em duas (ou mais) línguas que são traduções exatas um do outro. Já um corpus comparável contém textos sobre o mesmo assunto, mas que não são necessáriamente traduções um do outro. Existem ferramentas muito conhecidas e eficientes para trabalhar com corpora paralelos, que ajudam no desenvolvimento de dicionários e tradutores automáticos. Entretanto é muito difícil encontrar corpora paralelos, especialmente envolvendo o português. Por isso neste trabalho o foco é em técnicas de extração de termos aplicadas a corpora comparáveis, devido a sua maior disponibilidade e facilidade de se obtê-los com a ajuda da web.

O objetivo deste trabalho é aplicar técnicas de extração e alinhamento de termos multipalavras ao Português, produzindo assim recursos léxicos e avaliar a qualidade destes recursos utilizando diferentes técnicas e parametrizações. Para alcançar este objetivo foram implementados métodos de extração e alinhamento de termos, bem como ferramentas para auxiliar a avaliação da qualidade dos recursos produzidos. A avaliação de sucesso destas tarefas envolve a produção de uma lista de termos relevantes para o domínio do corpus escolhido, assim como a produção de uma lista ordenada de candidatos a termo correspondente na outra língua, onde o candidato correto encontra-se no topo da lista.

Utilizando corpora comparáveis para encontrar os termos alinhados, a principal metodologia empregada é a comparação do contexto dos termos, por isso temos como objetivo neste trabalho encontrar o tamanho do contexto ideal para o alinhamento e também descobrir quais as classes gramaticais mais adequadas para o contexto nesta tarefa.

Uma das principais motivações desse trabalho é criar recursos e ferramentas que possibilitem o avanço de diversas áreas do Processamento da Linguagem Natural, especialmente a Tradução Automática. Uma possível aplicação é na melhoria da tradução automática para a comunicação entre desenvolvedores de software falantes de línguas distintas, investigada por Calefato (2012). Os recursos construídos podem ser utilizados para alimentar o sistema de tradução automática melhorando a comunicação entre os desenvolvedores.

Outra motivação é investigar a aplicação dos métodos de alinhamento existentes na literatura para o Português, visto que sabemos que outras línguas, como Inglês, Francês e Alemão possuem mais recursos léxicos disponíveis. Assim queremos descobrir se esses métodos são aplicáveis ao Português e se podemos obter resultados equivalentes em nossa língua, mesmo ela tendo menos recursos disponíveis.

O restante desse trabalho será organizado da seguinte forma: no Capítulo 2 discutiremos sobre as características de um termo, no Capítulo 3 apresentaremos alguns trabalhos realizados

[2]Um corpus é uma coleção de textos organizada de alguma forma

na extração de termos, no Capítulo 4 descreveremos as ferramentas e materiais utilizados nesse trabalho, no Capítulo 5 descrevemos o sistema implementado, no Capítulo 6 mostramos os resultados que obtivemos e no Capítulo 7 concluímos.

2 FUNDAMENTAÇÃO TEÓRICA

Um dos maiores desafios do acesso multilíngue a informações de domínio específico é a detecção automática de terminologia. Para isso existem diversos estudos e abordagens, tanto sob uma perspectiva monolíngue, quanto multilíngue. Esse processo pode ser chamado de diversas formas: Reconhecimento de Termos, Identificação de Termos, Aquisição de Termos e Extração de Termos.

O resultado de um processo de extração de termos pode servir a diversos propósitos: construção de ontologias, índices para documentos em recuperação de informação, validação de memória de tradução. O comum entre todas as aplicações é o fato de que precisamos distinguir termos de não-termos (*computer science* vs. *additional information*), ou termos do domínio específico de vocabulário em geral.

O processo de Extração de Termos normalmente segue uma metologia padrão: leitura e pré-processamento do corpus, onde a entrada é uniformizada e anotada com informações linguísticas. Em seguida é feita a filtragem dos termos candidatos, onde todas as sequências de n palavras (n-gramas) são filtradas de acordo com algum critério linguístico definido pelo objetivo da extração. Por fim os candidatos que passaram pelo filtro são pesados utilizando alguma métrica, gerando uma lista ordenada de termos candidatos. Normalmente o processo automático de extração de termos é seguido por um processo manual de validação, por isso a saída de um processo de extração de termos é chamada de lista de termos candidatos. Para ajudar o processo de validação cada termo recebe um escore de *termhood*, e no caso multilíngue cada par <*termo, tradução candidata*> recebe um escore de similaridade.

Nas próximas seções apresentaremos algumas características básicas de um termo da perspectiva computacional, como as propriedades linguísticas (seção 2.1.1), estatísticas (seção 2.1.2) e distribucionais (seção 2.1.4). Também discutiremos como é avaliado o processo de extração monolíngue e multilíngue (seção 2.2). No restante do capítulo mostraremos alguns trabalhos importantes, primeiro para a extração monolíngue (seção 3.1) e por fim para extração multilíngue (seção 3.2).

2.1 Características dos Termos

Há um grande número de propriedades dos termos que os algoritmos de extração automática podem explorar, tanto durante a etapa de filtragem quanto na etapa de atribuir um escore a cada termo candidato, essas propriedades podem ser classificadas como linguísticas, estatísticas e distributivas.

2.1.1 Propriedades Linguísticas

A maioria dos trabalhos em extração de termos foca em sintagmas nominais (*noun phrases*, NP) ou grupos de substantivos (por exemplo "modelo conceitual", "teste unitário"), pois a maioria dos termos tendem a ser nominais, porém verbos e adjetivos também podem ser termos de domínios específicos (como em "manifesto ágil" e em "caso de uso"). Há também diversos trabalhos que focam na aquisição de contruções de um tipo específico de EM, como construções verbo-partícula (muito comuns no inglês, *take off, get on, make out*)(RAMISCH et al., 2008). Assim a estrutura linguística de termos é uma informação importante para uma extração automática de termos precisa para uma determinada língua. Por exemplo, para o inglês a estrutura foi descrita por Justeson e Katz (1995) como:

$$((Adj|Noun)^{+}|((Adj|Noun)^{*}(NounPrep)^{?}(Adj|Noun)^{*})Noun \tag{2.1}$$

Na definição 2.1, Adj representa adjetivos, Noun representa substantivos e Prep representa preposições. Esta definição inclui uma sequência de um ou mais adjetivos ou substantivos ($(Adj|Noun)^{+}$) ou uma sequência de adjetivos ou substantivos intercalados ou não por uma preposição ($(Adj|Noun)^{*}(NounPrep)^{?}(Adj|Noun)^{*}$), sempre terminados com um substantivo.

Para utilizar essa estrutura na extração de termos é preciso um corpus anotado com *part-of-speech* (PoS), ou categorias morfo-sintáticas de cada uma das palavras, i.e., substantivo, verbo, adjetivo, advérbio, entre outras. A estrutura línguistica captura muitos termos, porém captura muitos não-termos também. Por esse motivo, estas estruturas são um bom indicador de unicidade, mas não suficientemente restritas para serem um indicador de termo. Também foi observado que algumas palavras, mesmo sendo substantivos ou adjetivos raramente aparecem em termos, como por exemplo os adjetivos *following* e *interesting* e os subtantivos *thing* e *kind*. Essas palavras podem ser coletadas em uma lista de stopwords, e ignoradas na extração de termos.

Além da estrutura de *part-of-speech*, as relações entre diferentes partes da estrutura de um termo podem ser relevantes. Um termo complexo pode ser analisado a partir de seu componente principal com um ou mais modificadores (HIPPISLEY; CHENG; AHMAD, 2005). Em inglês o componente principal é geralmente o último substantivo não seguindo uma preposição, como em *database* ***manager*** e *flush* ***bolt*** *with fliplock*.

Outro aspecto linguístico é a estrutura morfológica. Em algumas línguas como sueco e alemão, palavras compostas são escritas como uma única palavra e é comum que essas palavras compostas sejam termos, portanto elas devem ser reconhecidas e preferencialmente sua estrutura interna seja identificada, ou seja, seu componente principal e seus modificadores. Como exemplos, podemos citar em sueco: katodstrålerör (tubo de raios catódicos), e em alemão: Prüfungstermin (data do exame)

Informações sobre radicais ou *lemas* das palavras são de uso limitado para extração de termos, mas são úteis para apresentação e validação, já que reduzem o número de candidatos que um validador deve inspecionar juntando todas as variantes de um termo em uma única expressão a ser validada (por exemplo: caso de teste, casos de teste, casos de testes). Assim como *part-of-speech*, o reconhecimento de radicais requer um pré-processamento usando ferramentas como lematizadores (SCHMID, 1994) e *stemmers* (ORENGO; HUYCK, 2001).

Além da estrutura interna das palavras e suas classes, os termos tem o seu contexto, que pode ser representado pelas palavras que ocorrem na vizinhança do termo ou por padrões típicos de frases em que os termos ocorrem, como definições e exemplificações. A vizinhança de um termo é representada por um vetor de palavras (dos quais trataremos bastante neste trabalho) e os padrões de ocorrência são capturados por expressões regulares como a seguinte (utilizada para o inglês), descrita por Pearson (1998):

$$(indef\_art)^{?} + term + connective\_verb + (def\_art|indef\_art)^{?} + (term|classword) + past\_participle \quad (2.2)$$

Na fórmula 2.2 temos um artigo indefinido (indef art) opcional, seguido do termo, de um verbo conetivo, mais um artigo definido (def art) ou indefinido opcional, e outro termo ou palavra representando a classe do termo, e um verbo no particípio passado. Uma instância dessa fórmula com *is* como verbo conetivo é "*a cucumber is a vegetable used...*".

2.1.2 Propriedades Estatísticas dos Termos

A propriedade estatítica básica de um termo é sua frequência no corpus bem como a de cada um de seus componentes individuais. Essas frequências podem ser usadas para a detecção de candidatos a termos e a comparação com a frequência em outros corpora, como um corpus balanceado ou um corpus de outro domínio para sua validação como termo específico de domínio.

Contagens básicas de frequência de uma palavra ou expressão em um corpus são usadas para computar medidas de co-ocorrência entre palavras.

2.1.2.1 Tabelas de Contingência

Uma estrutura muito usada para ajudar a entender e calcular as diversas medidas de associação para identificação de EM são as tabelas de contingência. Uma tabela de contingência (ver tabela 2.1) mostra todas as frequências relativas a um bigrama e suas ocorrências em um corpus.

As frequências nas células internas da tabela ($o_{11}, o_{12}, o_{21}, o_{22}$) são chamadas de frequências

Tabela 2.1 – Tabela de Contingência com as frequências observadas de um bigrama composto pelas palavras p_1 e p_2

	p_2	$\sim p_2$	
p_1	o_{11}	o_{12}	o_{1p}
$\sim p_1$	o_{21}	o_{22}	o_{2p}
	o_{p1}	o_{p2}	o_{pp}

observadas, pois denotam a quantidade de bigramas de cada tipo encontrado no corpus:

- o_{11} é o número de ocorrências do bigrama p_1p_2.
- o_{12} é o número de bigramas em que p_1 é a palavra da esquerda e p_2 não é a palavra da direita.
- o_{21} é o número de bigramas em que p_1 não é a palavra da esquerda e p_2 é a palavra da direita.
- o_{22} é o número de bigramas em que p_1 não é a palavra da esquerda e p_2 não é a palavra da direita.

Somando todas as frequências observadas temos o número total de bigramas no corpus o_{pp}. Os totais nas linhas (o_{1p}, o_{2p}) e nas colunas (o_{p1}, o_{p2}) são chamados de frequências marginais e representam a soma das frequências observadas nas linhas e colunas:

- o_{1p} é o número de ocorrências de p_1 à esquerda de um bigrama, também é igual a frequência de p_1 no corpus.
- o_{2p} é o número de bigramas em que a palavra à esquerda não é p_1.
- o_{p1} é o número de ocorrências de p_2 à direita de um bigrama, também é igual a frequência de p_2 no corpus.
- o_{p2} é o número de bigramas em que a palavra à direita não é p_2.

Onde 1 significa a ocorrência da palavra na posição indicada, 2 significa a não ocorrência da palavra naquela posição e p significa a ocorrência de qualquer palavra.

Tabelas de contingência para trigramas não são muito utilizadas pois são tridimensionais(ANDERSEN, 2012). Assim, em uma tabela de contingência para um trigrama teremos as respectivas frequências observadas:

- o_{111} é o número de ocorrências do trigrama $p_1p_2p_3$.
- o_{112} é o número de trigramas em que p_1 é a palavra da esquerda, p_2 é a palavra do meio e p_3 não é a palavra da direita.
- o_{121} é o número de trigramas em que p_1 é a palavra da esquerda, p_2 não é a palavra do meio e p_3 é a palavra da direita.
- o_{122} é o número de trigramas em que p_1 é a palavra da esquerda, p_2 não é a palavra do meio e p_3 não é a palavra da direita.

- o_{211} é o número de trigramas em que p_1 não é a palavra da esquerda, p_2 é a palavra do meio e p_3 é a palavra da direita.
- o_{212} é o número de trigramas em que p_1 não é a palavra da esquerda, p_2 é a palavra do meio e p_3 não é a palavra da direita.
- o_{221} é o número de trigramas em que p_1 não é a palavra da esquerda, p_2 não é a palavra do meio e p_3 é a palavra da direita.
- o_{222} é o número de trigramas em que p_1 não é a palavra da esquerda, p_2 não é a palavra do meio e p_3 não é a palavra da direita.

As frequências marginais mais utilizadas para trigramas são:

- o_{1pp} é o número de ocorrências de p_1 à esquerda de um trigrama, também é igual a frequência de p_1 no corpus.
- o_{p1p} é o número de ocorrências de p_2 no meio de um trigrama, também é igual a frequência de p_2 no corpus.
- o_{pp1} é o número de ocorrências de p_3 à direita de um trigrama, também é igual a frequência de p_3 no corpus.
- o_{2pp} é o número de trigramas em que a palavra à esquerda não é p_1.
- o_{p2p} é o número de trigramas em que a palavra do meio não é p_2.
- o_{pp2} é o número de trigramas em que a palavra à direita não é p_3.

2.1.2.2 Modelo de Independência

Muitas medidas de associação tentam comparar uma tabela de contingência de valores observados com uma de frequências esperadas. As frequências esperadas de uma tabela de contingência são estimadas baseada em um modelo de independência. A hipótese desse modelo é de que as palavras em um n-grama co-ocorrem ao acaso (são independentes). Para bigramas, esta hipótese significa que a probabilidade das duas palavras ocorrerem juntas é igual ao produto das suas probabilidades individuais de ocorrência (equação 2.3). Sendo que a probabilidade de uma palavra $p1$ ocorrer em um texto com N palavras é dada pela equação 2.4,

$$P(p1, p2) = P(p1) * P(p2) \tag{2.3}$$

$$P(p1) = \frac{o_{1p}}{N} \tag{2.4}$$

Usando esta hipótese podemos calcular a tabela de frequências esperadas usando os totais marginais e o total de n-gramas, visto na tabela 2.2.

Como exemplo de uma tabela de contingência temos o número de ocorrências do bigrama Sherlock Holmes na novela *A Case of Identity* ilustrado na tabela 2.3.

Tabela 2.2 – Tabela de Contingência com as frequências esperadas de um bigrama composto pelas palavras p_1 e p_2

	p_2	$\sim p_2$	
p_1	$e_{11} = \frac{n_{p1} * n_{1p}}{n_{pp}}$	$e_{12} = \frac{n_{p2} * n_{1p}}{n_{pp}}$	n_{1p}
$\sim p_1$	$e_{21} = \frac{n_{p1} * n_{2p}}{n_{pp}}$	$e_{11} = \frac{n_{p1} * n_{1p}}{n_{pp}}$	n_{2p}
	n_{p1}	n_{p2}	n_{pp}

Tabela 2.3 – Frequências observadas do bigrama *Sherlock Holmes*

	holmes	**~holmes**	
sherlock	7	0	7
~sherlock	39	7059	7098
	46	7059	7105

A partir das frequências marginais podemos calcular as frequências esperadas, vistas na tabela 2.4.

Tabela 2.4 – Frequências esperadas do bigrama *Sherlock Holmes*

	holmes	**~holmes**	
sherlock	0,05	6,95	7
~sherlock	45,95	7052,05	7098
	46	7059	7105

2.1.3 Medidas de Associação

A partir das frequências observadas e das frequências esperadas para uma determinada EM, podemos calcular diversas medidas de associação usadas para identificação de EM, entre elas: *Log Likelihood Ratio, Poisson Stirling, Pointwise Mutual Information, True Mutual Information* e o coeficiente de *Dice*.

2.1.3.1 Log Likelihood Ratio

O **Log Likelihood Ratio** (LL, equação 2.5) proposto por Wilks (1938), e aplicado a n-gramas por Dunning (1993), verifica a probabilidade de amostrar uma tabela de contingência observada sobre a hipótese nula (de que as palavras são independentes). Ele mede a diferença entre os valores observados e os valores esperados. É soma da razão entre os valores observados e os valores esperados.

$$LL = 2 * \sum o_{ij} * log\left(\frac{o_{ij}}{e_{ij}}\right) \tag{2.5}$$

2.1.3.2 *Poisson Stirling*

A medida **Poisson Stirling** (PS, equação 2.6) teve sua aplicação à identificação de EM proposta por (QUASTHOFF; WOLFF, 2002), ela é uma medida de *likelihood*, que procura estimar a probabilidade de termos uma tabela de contingência observada contra a hipótese nula de que as palavras no n-grama ocorreram juntas ao acaso e são independentes.

$$PS = o_{11} * (log\left(\frac{o_{ij}}{e_{ij}}\right) - 1) \tag{2.6}$$

2.1.3.3 *Pointwise Mutual Information*

Outra medida muito usada é o *Pointwise Mutual Information* ou PMI (CHURCH; HANKS, 1990) definida na equação 2.7, que compara a probabilidade das duas palavras aparecerem juntas se elas forem independentes ($P(p_1)P(p_2)$) com a probabilidade atual delas aparecerem juntas ($P(p_1p_2)$).

Como a probabilidade de ocorrência de p_1p_2 pode ser expressa por $P(p_1p_2) = \frac{n_{11}}{n_{pp}}$, assim como as probabilidades das palavras ocorreram sozinhas são $P(p_1) = \frac{n_{1p}}{n_{pp}}$ e $P(p_2) = \frac{n_{p1}}{n_{pp}}$, e também sabendo que a frequência esperada para p_1p_2 é dada por $e_{11} = \frac{n_{p1} * n_{1p}}{n_{pp}}$, podemos substituir essas equações na equação 2.7 e obter uma fórmula mais simples para o pmi na equação 2.8.

$$PMI(p1p2) = \log \frac{P(p_1p_2)}{P(p_1)P(p_2)} \tag{2.7}$$

$$PMI = log\frac{o_{11}}{e_{11}} \tag{2.8}$$

2.1.3.4 *True Mutual Information*

True Mutual Information (TMI, equação 2.9) é muito parecida com Log-Likelihood, a diferença é que ao invés de somente multiplicar a razão dos valores observados pelos valores esperados pela frequência observada (ex: n_{11}), eles são multiplicados pela razão entre a frequência observada e o total de bigramas (ex: $\frac{n_{11}}{n_{pp}}$).

$$TMI = \sum \frac{o_{ij}}{o_{pp}} * log\left(\frac{o_{ij}}{e_{ij}}\right) \tag{2.9}$$

2.1.3.5 *Coeficiente Dice*

Outra medida de associação é o coeficiente *Dice* (equação 2.10), usado para extração de EM pela primeira vez por Smadja (1993), onde um valor alto do coeficiente de Dice significa que o

as palavras não ocorrem juntas ao acaso e são um bom candidato a EM.

$$Dice = \frac{2 * o_{11}}{o_{1p} + o_{p1}} \tag{2.10}$$

2.1.3.6 Outras Frequências e Medidas

Outra informação relevante sobre frequência diz respeito a termos aninhados, quando um candidato a termo sobrepõe outros candidatos mais longos ou mais curtos. Essa situação é muito comum como por exemplo as expressões *floating point* e *floating point arithmetic* que são ambas termos, enquanto *point arithmetic* não é. Para esses casos as seguintes frequências são relevantes (Frantzi (1998) e Basili et al (2001)):

1. A frequência de um termo candidato contido em outro candidato.
2. A frequência de um termo candidato como modificador ou componente principal do termo.
3. O número de candidatos mais longos dos quais o termo em questão faz parte.
4. O tamanho, em número de palavras, do termo.

Além destas medidas apresentadas, existem muitas outras, como podemos ver no trabalho de Pecina (2010), que investiga mais de 80 medidas de associação usadas para identificação de expressões multipalavra. Também foram investigadas técnicas de classificação baseadas em aprendizado de máquina, como redes neurais e máquinas de vetor de suporte e em todos os testes realizados a combinação de múltiplas medidas de associação foi melhor que a utilização delas individualmente.

2.1.4 Propriedades Distribucionais dos Termos

As propriedades distribucionais dos termos podem ser vistas como casos particulares das estatísticas, já que são baseadas em contagens. Elas podem ser relacionadas à distribuição dos termos em um corpus (prevalência) e também à comparação da distribuição em um corpus de domínio específico em relação à distribuição em um corpus geral. Uma medida muito comum, muito utilizada na área de Recuperação de Informação é o *tf-idf*, onde *tf* (*term frequency*) é a frequência de um termo em um documento e *idf* é a frequência inversa do documento, calculada da seguinte forma para o termo t_i:

$$IDF = \log \frac{N}{n_i} \tag{2.11}$$

Onde N é o número de documentos no corpus e n_i é o número de documentos onde t_i ocorre, e a medida final *tf-idf* é dada pela multiplicação do *tf* pelo *idf*. A idéia dessa medida é que palavras muito comuns (que aparecem em muitos documentos) tem um *idf* baixo e não tem tanta importância quanto palavras mais raras, mesmo que elas tenham uma frequência alta.

Outra medida que compara a distribuição de um termo em dois corpus, um deles específico e outro geral, é a *Weirdness*, descrita na equação 2.12 onde usamos D para um corpus de domínio específico, G para um corpus geral, f para frequência e N para o tamanho do corpus.

$$Weirdness = \frac{f_D * N_G}{f_G * N_D} \tag{2.12}$$

2.2 Avaliação

A avaliação de sistemas de extração de termos é um problema difícil, já que a definição de um termo depende de critérios externos. *Gold Standards* são difíceis de serem encontrados, assim dicionários, ontologias e juízes humanos são frequentemente utilizados para avaliação.

A avaliação é focada na precisão, principalmente quando juízes humanos são utilizados, pois seria trabalhoso identificar todos os termos em um corpus de teste. Também é raro utilizar toda a saída do sistema na forma de lista de candidatos possivelmente ranqueada por alguma medida, geralmente a avaliação é restrita a uma parte da saída. A precisão é medida como sendo a proporção ou percentual de candidatos avaliados que são considerados termos (equação 2.13).

$$P = \frac{numeroDeTermos}{numeroDeCandidatos} \tag{2.13}$$

Também é comum calcular a precisão de acordo com um intervalo de frequência. Outra medida alternativa, que também captura a revocação, dado um *gold standard* é a UAP (*Uninterpolated Average Precision*) definida na equação 2.14

$$UAP = \frac{1}{K} \sum_{i=1}^{K} P_i \tag{2.14}$$

Onde P_i representa a "precisão em i", e i é um número de termos corretos. P_i é computado como a taxa $\frac{i}{H_i}$ onde H_i é o número de candidatos necessários para se obter i termos corretos. É comum que P_i diminua conforme i aumenta, pois quanto mais candidatos avaliamos, a tendência é encontrarmos mais candidatos incorretos, mas o UAP considera uma média de todos os i considerados. Assim, dada uma lista ranqueada se espera que no topo da lista estejam os candidatos corretos e um valor alto de UAP.

Outra medida utilizada é o ruído (*Noise*), o dual da Precisão, que mede o percentual de termos candidatos não-utilizáveis, também chamados de falsos positivos, ao invés dos úteis.

Os resultados de avaliações podem variar de acordo com o corpus e os métodos de avaliação utilizados, para corpus extraídos da Wikipedia Hjelm (2009) obteve valores de precisão baixos, entre 12-13% enquanto Zhang, Brwster e Ciravegna (2008) encontraram valores por volta de 90% também na Wikipedia.

Outro fato importante sobre a avaliação de extração de termos, é que muitos trabalhos utilizam métodos de avaliação diferentes dos usuais, avaliando por vezes algum aspecto que seu

trabalho focou, assim dificultando a comparação entre os métodos existentes na literatura.

3 TRABALHOS RELACIONADOS

Nesse capítulo apresentamos alguns trabalhos que serviram como base para este. Primeiro falaremos sobre algumas métricas propostas para identificação de expressões multipalavra e depois de técnicas para o alinhamento de expressões multipalavra multilíngue.

3.1 Extração de Termos Monolíngue

3.1.1 O método c-value/nc-value

O método chamado *c-value/nc-value* (FRANTZI; ANANIADOU; TSUJII, 1998) é uma combinação mais elaborada de métodos linguísticos e estatísticos. O *c-value* é calculado a partir da frequência f_a de um candidato a, seu tamanho (em número de palavras) $|a|$ e do conjunto de termos T_a que contém a.

$$\text{c-value} = \log_2 |a| f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b \tag{3.1}$$

A idéia é subtrair do *score* básico, calculado a partir da frequência e do número de termos do candidato a, a frequência média dos termos que contém a. É importante notar que caso o termo candidato a não seja contido por outro, o segundo termo da equação será zero. O *c-value* deve ser aplicado em termos candidatos que passarem por um filtro linguístico baseado em *part-of-speech* e uma lista de *stopwords*. Esse filtro pode ser balanceado para privilegiar precisão ou *recall*. O *c-value* vai ter um valor alto para termos longos e não contidos em outros que tiverem uma frequência alta, por outro lado candidatos não maximais, que fazem parte de candidatos com alta frequência terão um *c-value* baixo.

O *c-value* foi avaliado com um corpus médico utilizando precisão, e comparado com a extração de termos utilizando somente a frequência, os resultados obtidos por Frantzi, Ananiadou e Tsujii (1998) foram melhores em todos os três casos avaliados: termos que apareceram somente aninhados em outros, termos que apareceram também contidos em outros e todos os termos. Nos dois primeiros casos as melhoras foram expressivas, chegando a 38%, já no último caso, não houveram melhoras significativas, já que o método trata termos que não aparecem contidos em outros praticamente da mesma forma que a frequência.

O cálculo do *c-value* é o primeiro passo do método nc-value, no segundo passo identificamos um conjunto de *palavras de contexto do termo* e associamos a cada uma delas um peso. A idéia é que um termo é caracterizado por, além de suas próprias características como a frequência, o contexto em que ele ocorre, nesse caso, as palavras que ocorrem junto com o termo. Essas *palavras de contexto do termo* são verbos, adjetivos e substantivos que ocorrem na vizinhança dos termos, o peso associado a cada uma dessas palavras busca medir sua importância por meio do número de termos que aparece junto com a palavra. O peso de uma palavra de contexto é

definido na equação 3.2:

$$weight(w) = \frac{t(w)}{T} \tag{3.2}$$

Onde $t(w)$ é o número de termos que aparecem junto com a palavra w e T é o número de termos considerados. Essa medida representa o percentual de termos em que a palavra de contexto apareceu em conjunto.

A avaliação dessa metodologia de extração de palavras de contexto foi feita por Frantzi, Ananiadou e Tsujii (**??**) da seguite maneira: uma lista de termos foi extraída de um corpus usando *c-Value* e manualmente revisada, a partir desses termos, uma lista de palavras de contexto foi montada utilizando a equação 3.2. Dessa lista de palavras de contexto foram extraídos três conjuntos de 20 palavras cada: um do topo da lista, um do meio e outro da parte inferior. Foram contados os números de termos que ocorrem com cada uma das palavras de contexto de cada um dos conjuntos. Somando os números de cada conjunto, os resultados obtidos foram que as palavras do topo da lista estão associadas a 12% mais termos que as palavras do meio, que por sua vez estão associadas a 21% mais termos que as palavras da parte inferior da lista. Dessa maneira foi demonstrado que essa medida cumpre seu propósito de associar valores mais altos a palavras que ocorrem mais junto com termos.

Depois de extraídas, as **palavras de contexto dos termos** são usadas para calcular o *nc-value*, que consiste em três etapas:

1. Aplicar o *c-value* ao corpus, como resultado é obtida uma lista de termos ordenada pelo *c-value*.
2. A segunda etapa é a extração de **palavras de contexto**, que necessita de uma lista de termos. Nessa etapa os autores ressaltam que resolveram utilizar os primeiros 5% candidatos gerados pela etapa anterior para manter o método totalmente automático e sem a utilização de nenhuma fonte externa, como dicionários, para esta etapa.
3. A terceira etapa consiste em incorporar a informação de contexto adquirida na segunda etapa. A lista de termos gerada pelo *c-Value* é re-ranqueada para que os verdadeiros termos apareçam com maior concentração no topo. Para cada termo candidato é calculado um fator de contexto somando os pesos de suas palavras de contexto multiplicados pela frequência que a palavra aparece junto com o termo. Esse é o segundo fator na equação 3.3 que descreve o método *nc-value*, onde C_a é o conjunto das palavras de contexto do termo a, f_ab é a frequência que a palavra de contexto b aparece junto com o termo a e $weight(b)$ é peso da palavra de contexto b. Os fatores $0,8$ e $0,2$ foram escolhidos pelos autores com base em experimentos realizados.

$$\text{nc-value} = 0,8 * c - value(a) + 0,2 * \sum_{b \in C_a} f_ab * weight(b) \tag{3.3}$$

A metodologia *nc-value* foi avaliada no mesmo corpus que o *c-value* e apresentou uma melhora de 5% na concentração de termos no topo da lista de termos extraídos.

3.1.2 *Contrastive Selection via Heads*

O *Contrastive Selection via Heads* (BASILI et al., 2001) é uma medida que busca levar em conta a distribuição dos candidatos a termos em um corpus de domínio e em um (ou mais) corpus fora do domínio. O primeiro passo do método é a extração de candidatos com a utilização de um parser para determinar a estrutura do termo e encontrar o componente principal do termo. Os termos extraídos são divididos em simples (os componentes principais dos termos extraídos) e complexos (termos completos, constituídos de componente principal e modificadores). Após a extração dos termos e seus componentes principais, é aplicado um filtro estatístico.

Aos termos simples é atribuído um *score* utilizando Frequência Inversa da Palavra (em inglês, *Inverse Word Frequency*, IWF) uma adaptação de Frequência Inversa do Documento (em inglês, *inverse document frequency*, IDF) para extração de termos, descrita na equação 3.4, onde N é a soma das frequências de todos os candidatos em todos os corpus e $F_t = \sum_j f_t^j$, com j sendo cada um dos corpus utilizados, ou seja, F_t é a soma das frequências de t em todos os corpus.

$$IWF(t) = log(\frac{N}{F_t}) \tag{3.4}$$

Como o IWT não leva em conta o domínio, o *score* final do termo é calculado multiplicando o IWF pela frequência do termo no domínio, segundo a equação 3.5, onde i é o domínio em questão e t o termo.

$$w_t^i = log(f_t^i) * IWF(t) \tag{3.5}$$

Assim obtemos o *contrastive weight*, w_t^i dos termos simples (componentes principais), para obter o valor final de um termo complexo ct no corpus i (cw_{ct}^i), é multiplicado o w_t^i dele, pela sua frequência no corpus i, como visto na equação 3.6.

$$cw_{ct}^i = w_{h(ct)}^i * f_{ct}^i \tag{3.6}$$

Os resultados do método proposto (BASILI et al., 2001) foram avaliados com um corpus do Código Civil Italiano, como corpus constrastivo foi utilizado um corpus de notícias de domínios variados (esportes, política e economia). Como referência uma lista de 2000 termos foi manualmente construída e validada, porém somente 943 desses termos estão presentes no corpus, esses termos, chamados de sobrepostos foram utilizados como *gold standard* para calcular a *f-measure*, utilizada como métrica para comparar a medida proposta e a frequência do termo no corpus de domínio.

A *f-measure* foi comparada para diferentes números de termos selecionados e os resultados da medida proposta, o *Constrastive Weight* foi sempre melhor que a frequência pura.

Um dos fatores destacados pelos autores é que a avaliação de um sistema extrator de termos é uma tarefa muito difícil, pois gostaríamos que o *gold standard* utilizado fosse altamente controlado e validado por especialistas, porém muitos termos sugeridos pelo sistema, considerados

termos de domínio por não-especialistas não estão presentes no *gold standard*. Desse modo uma avaliação baseada em um *gold standard* desse tipo é inadequada para capturar totalmente a natureza da tarefa.

3.1.3 *Discriminative Weight*

A metodologia de extração de termos proposta por Wong, Liu e Bennamoun (2007) busca explorar quatro aspectos que não são bem tratados nos trabalhos da área:

1. Atenção inadequada à diferença entre prevalência e tendência: Muitas técnicas existentes baseadas em frequência são adaptadas do tf-idf, mas essas medidas não refletem a tendência de utilização dos termos através de diferentes domínios mas simplesmente medem a prevalência do termo em um domínio específico.
2. Simplificação do papel de componentes principais e modificadores: Muitas abordagens tentaram utilizar o componente principal como representativo de todo o termo complexo. Por exemplo, a afirmação "o sentido do termo é geralmente determinado por seu componente principal" (BASILI et al., 2001), não é completamente verdadeira, já que os componentes principais são inerentemente ambíguos e os modificadores são necessários para diminuir suas possíveis interpretações.
3. Como determinar o relacionamento entre termos e seu contexto? Para abordagens que usam informação contextual ainda não é conhecida a melhor maneira de explorá-las. Por exemplo, Maynard e Ananiadou (1999) usam recursos raros e estáticos para a computação de similaridade, enquanto Basili et al (2001) usa grandes corpora para extração de contexto como características.
4. A ênfase nas evidências contextuais: Muitos pesquisadores repetem o clichê: "você conhece uma palavra pela companhia dela", mas a menos que você tenha os mecanismos para identificar as companias que realmente identificam a palavra a ênfase nas evidências contextuais pode gerar resultados negativos.

O mecanismo proposto neste trabalho (WONG; LIU; BENNAMOUN, 2007) utiliza dois tipos de corpus: um corpus de domínio d e um corpus contrastivo $\overline{d}$. Abaixo estão descritas duas medidas básicas utilizadas para capturar as evidências estatísticas baseadas na distribuição entre domínios e na distribuição intra-domínio:

1. DP (*Domain Prevalence* ou prevalência do domínio), que mede o grau de utilização do termo no corpus do domínio de destino. A distribuição intra-domínio dos termos candidatos e das palavras de contexto são utilizadas para calcular a DP.
2. DT (*Domain Tendency* ou tendência do domínio), que mede o grau de inclinação da utilização do termo para o domínio de destino. O comportamento distributivo entre domínios dos termos candidatos e palavras de contexto são usados para calcular a DT.

Um alto DP significa que o termo é frequente no domínio. Ele é um sinal de relevância para o domínio se e somente se a frequência de uso do termo é exclusiva do domínio, ou seja, também há uma alta DT.

Abaixo estão os três tipo de evidências linguísticas utilizadas:

1. DW (*Discriminative Weight* ou peso discriminativo) é o produto da tendência do domínio e da prevalência do domínio. Essa evidência é o primeiro passo para isolar os candidatos relevantes ao domínio dos candidatos gerais.
2. MF (*Modifier Factor* ou fator modificador) é obtido computando a DT usando a frequência cumulativa dos modificadores de termos complexos.
3. ACDW (*Average Contextual Discriminative Weight* ou peso discriminativo contextual médio) é calculada usando a DW cumulativa das palavras de contexto, escaladas de acordo com com sua relação semântica com os termos candidatos correspondentes. ACDW é ajustado utilizando a DW do termo candidato para obter a ACC (*Adjusted Contextual Contribution* ou contribuição contextual ajustada) para refletir a confiança das evidências contextuais.

A partir de uma lista de termos candidatos, $TC = \{a_1, a_2, ..., a_n\}$, o objetivo é atribuir scores para ranquear os termos na lista. Cada candidato pode ser simples (somente elemento principal), ou complexo (contendo elemento principal e modificadores).

O DP calculado é fortemente baseado na abordagem Constrastive Weight (BASILI et al., 2001), e é calculado de uma forma para termos simples e de outra para termos complexos. Se a é um termo simples, a DP é calculada de acordo com a fórmula 3.7

$$DP(a) = log_{10}(f_{ad} + 10) log_{10}(\frac{F_{TC}}{f_{ad} + f_{a\bar{d}}} + 10) \tag{3.7}$$

Onde F_{TC} é a soma das frequências de ocorrência de todos os termos $a \in TC$ no corpus de domínio e no corpus contrastivo, f_{ad} é a frequência do termo a no corpus de domínio e $f_{a\bar{d}}$ é a frequência do termo a no corpus contrativo. Se o termo a é um termo complexo, DP é definida de acordo com a equação 3.8:

$$DP(a) = log_{10}(f_{ad} + 10) DP(a^h) MF(a) \tag{3.8}$$

Assim como no *Contrastive Weight* original (BASILI et al., 2001), é usado o componente principal para calcular a DP para termos complexos. No entanto, foi notado que a multiplicação direta de f_{ad} de termos complexos muito comuns causa uma distorção nos pesos e dá uma falsa impressão de importância no domínio, por isso foi adicionado a constante 10 e o logaritmo na frequência dos termos complexos.

Além disso a principal contribuição deste método está no MF, que tem como função atribuir pesos maiores do que seus componentes principais a candidatos complexos relevantes. Esse

fator também é usado para penalizar termos complexos que tenham elementos principais relacionados ao domínio, mas não são do domínio, como, por exemplo, o elemento principal "virus" no domínio "tecnologia" tem um peso alto, porém se ele for moficado por "H5N1", formando "H5N1 virus", o termo não é importante para o domínio "tecnologia". Da mesma forma esse fator é usado para compensar baixos pesos recebidos por termos complexos relevantes que possuem elementos principais não relacionados ao domínio, como "account" que não está relacionado ao domínio tecnologia, mas se ele vier modificado por "Google" ou "Gmail" será relevante para esse domínio.

O fator MF de um termo complexo a é medido usando a frequência cumulativa de domínio e a frequência cumulativa contrastiva dos modificadores que também são termos candidatos, $m \in M_a \cap TC$. Formalmente o MF de um termo complexo é definido na equação 3.9:

$$MF(a) = log_2(\frac{\sum_{m \in M_a \cap TC} f_{md} + 1}{\sum_{m \in M_a \cap TC} f_{m\bar{d}} + 1} + 1) \tag{3.9}$$

MF é derivado da DT, tendência do domínio, a única diferença entre entre elas é que MF utiliza os modificadores enquanto a DT utiliza o termo candidato completo, simples ou complexo. MF e DT são duas medidas discriminantes que ajudam a diferenciar candidatos que são relevantes no domínio dos candidatos que são de uso geral. Formalmente a DT de um candidato a é calculada segundo a equação 3.10:

$$DT(a) = log_2(\frac{f_{ad} + 1}{f_{a\bar{d}} + 1} + 1) \tag{3.10}$$

Se o termo candidato ocorrer igualmente no corpus de domínio e no corpus contrastivo, $DT = 1$, se o uso dele for maior no corpus de domínio, $DT > 1$ e se o uso for maior no corpus contrastivo, $DT < 1$. As duas medidas básicas DT e DP juntas formam um novo peso, chamado de DW, descrito na equação 3.11:

$$DW(a) = DP(a)DT(a) \tag{3.11}$$

Assumindo que um termo candidato a tem um conjunto de palavras de contexto C_a, o ACDW (*average contextual discriminative weight*) é definido pela equação3.12:

$$ACDW(a) = \frac{\sum_{c \in C_a} DW(C) sim(a, c)}{|C_a|} \tag{3.12}$$

Onde $sim(a, c) = 1 - NGD(a, c)\theta$, $NGD(a, c)$ é a distância normalizada do Google (Normalized Google Distance) (CILIBRASI; VITANYI, 2007) entre o termo candidato a e a palavra de contexto c, e θ é uma constante para escalar o valor da NGD. O ACDW é usado para levar em consideração o contexto dos termos, mas nesse trabalho o ACDW não tem uma contribuição direta para o peso final de um termo. Duas medidas são utilizadas para ajustar a contribuição do

peso do contexto, primeiro o NGD, que quantifica a relação entre um termo e sua palavra de contexto, usado no cálculo do ACDW para que palavras de contexto mais relacionadas ao termo tenham uma contribuição maior para o ACDW. Segundo, o ACDW é ajustado de acordo com sua taxa com o correspondente DW para produzir o ACC (*Adjusted Contextual Contribution*), formalmente definido na equação 3.13:

$$ACC(a) = ACDW(a)\frac{e^{(1-\frac{ACDW(a)+1}{DW(a)+1})}e^{(1-\frac{DW(a)+1}{ACDW(a)+1})}}{log_2\frac{ACDW(a)+1}{DW(a)+1}+1} \tag{3.13}$$

O peso final de um termo, chamado de Termhood (TH) é calculado pela equação 3.14:

$$TH(a) = DW(a) + ACC(a) \tag{3.14}$$

Os experimentos foram realizados com um corpus de domínio com 24 documentos (51.289 palavras) sobre "*liver cancer*" (cancer de fígado) extraído do BioMedCentral.com e um corpus contrastivo de 11.115 artigos (4.378.210 palavras) de vários assuntos extraídos de Reuters.com, CNet.com e ABS.com. Os resultados da métrica proposta na equação 3.14 foram comparados com Contrastive Weight (BASILI et al., 2001) e com nc-value (FRANTZI; ANANIADOU; TSUJII, 1998).

Os resultados dos três métodos foram comparados em termos da distribuição das frequências dos termos no corpus de domínio e no corpus contrastivo. Os termos com maior valor de TH foram aqueles com uma alta frequência no corpus de domínio e com uma baixa frequência no corpus contrastivo, os termos com um alta frequência no corpus contrastivo receberam valores baixos de TH.

3.2 Extração de Termos Multilíngue

A extração de termos multilígue é a extração de termos de uma ou mais línguas com o objetivo de criar ou estender um recurso existente como um dicionário bilíngue, um glossário, ou uma ontologia. A diferença geral entre a extração monolígue e multilíngue é que no caso multilíngue geralmente não queremos somente extrair os termos, mas associar os termos correspondentes entre as línguas. Para extração de termos multilíngue precisamos de documentos nas línguas de interesse. Caso esses documentos sejam originais e traduções temos um **corpus paralelo**, se eles tiverem assuntos relacionados, temos um **corpus comparável**.

O processo mais simples de extração multilíngue, chamado de *term spotting* ou *translation spotting* assume que temos um corpus paralelo e conhecemos a lista de termos em uma língua e queremos encontrar os equivalentes na outra. Esse processo pode ser feito para qualquer unidade linguística, desde uma palavra até uma frase inteira. O processo mais comum utilizado é o *word alignment* ou alinhamento de palavras, o sistema mais utilizado nesse processo é o Giza++ (OCH; NEY, 2003) que usa o algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977)

aplicado ao alinhamento de palavras e tradução automática (JURAFSKY; MARTIN, 2008).

Primeiro iremos apresentar um trabalho que usa o Giza++ e corpora paralelos, e em seguida iremos analisar os métodos que podem ser usados para extração multilíngue de termos multipalavra em corpora comparáveis.

3.2.1 Extração de expressões multipalavra de pequenos corpora paralelos

O trabalho de Tsvetkov e Wintner (2012) propõe um novo algoritmo para identificação de expressões multipalavra (*multiword expressions*, ou MWEs) em corpus bilíngue, usando *automatic word alignment* como fonte primária de informação. Diferente das abordagens existentes, a busca não é limitada a alinhamentos um-para-muitos, e propõe uma estratégia de mineração de erros para detectar os desalinhamentos no corpus paralelo. Também é consultado um grande corpus monolíngue para ordenar e filtrar as expressões. O resultado é a extração totalmente automática de expressões multipalavra de vários tipos, tamanhos e padrões sintáticos, junto com suas traduções. A utilidade dessa metodologia é demonstrada incorporando o dicionário extraído em um sistema de tradução automática existente.

A língua de origem utilizada é o hebraico e o corpus paralelo hebraico-inglês tem 19.626 frases, extraídas em sua maioria de jornais. Esse trabalho é focado em expressões multipalavra não composicionais, isto é, aquelas em que a tradução das partes não é igual a tradução do todo, por exemplo, a expressão *foreign worker*, em hebraico *wbd zr* , que é traduzida diretamente não será identificada.

Devido às características do hebraico, e às grandes diferenças do inglês, o pré-processamento é uma etapa muito importante, nele são feitas a tokenização e lematização do corpus. A importância dessa etapa se deve ao fato de que em hebraico informações de gênero, número, pessoa e modo se refletem na morfologia da palavra, além disso preposições, artigos e conjunções podem estar concatenados às palavras como prefixos ou sufixos.

O alinhamento do corpus paralelo, após o pré-processamento é feito pelo Giza++. As expressões multipalavra são buscadas no corpus paralelo procurando por desalinhamentos, ou seja, são analisadas todas as palavras que não foram alinhadas como 1:1, além disso todas as palavras alinhadas como 1:1 são verificadas por um dicionário bilíngue. Um corpus monolíngue é usado para calcular as estatísticas da distribuição da sequência de palavras em hebraico. Corpus monolíngues são mais fáceis de obter e produzem estatíticas mais confiáveis. Ele é usado para validar as MWE's candidatas produzidas pela técnica baseada em alinhamento, por exemplo, expressões com baixa frequência são eliminadas.

É usada uma variação de PMI, o PMI^k após os alinhamentos 1:1 serem removidos para associar um score a cada bigrama extraído dos desalinhamentos. É determinado um ponto de corte, e todas as sequências de palavras em que todos os bigramas tem um score maior que o ponto de corte são consideradas termos multipalavra.

Os resultados encontrados são expressões multipalavra de diversos tipos: nomes próprios,

locais geográficos, compostos nominais, combinações de adjetivos e substantivos, entre outros. Os 100 primeiros candidatos foram avaliados por 2 anotadores e 99 foram considerados MWEs.

Diversas avaliações foram feitas:

1. Usando uma lista extraída de um corpus anotado com 201 exemplos positivos de MWEs e 91 exemplos negativos de compostos nominais, o algoritmo proposto foi comparado com 4 baselines, e foi o que apresentou melhor f-score, com o melhor *recall* entre todos.
2. Para demonstrar a importância do pré-processamento, foi realizada uma comparação entre o método proposto com e sem pré-processamento. Os resultados foram que o número de alinhamentos 1:1 encontrados no dicionário (que podem ser descartadas) pelo método sem pré-processamento foi de 10% em comparação com o método com pré-processamento, o que torna a tarefa muito mais difícil. Como resultado final, grande parte das expressões multipalavra encontradas pelo método sem pré-processamento foram nomes próprios.
3. A última avaliação foi feita treinando um tradutor automático utilizando o MOSES (KOEHN et al., 2007) com diferentes configurações, e utilizando o score BLEU (PAPINENI et al., 2002) para avaliar a qualidade das traduções. O corpus paralelo com aproximadamente 20.000 frases foi usado para treinar o modelo de tradução, e um grande corpus monolíngue de jornais em Inglês foi usado para treinar o modelo de linguagem. A partir desse tradutor baseline foram treinados outros 3: o 1º utilizando somente as 1.000 melhores MWEs (e suas traduções) encontradas, o 2º utilizando todas as MWEs (3.750), e o 3º utilizando todas, mas colocando-as três vezes no corpus para aumentar seu peso. Em todos os casos, o score BLEU foi melhor que o do baseline, e o melhor de todos foi o 2º, que obteve uma melhora significativa em relação ao baseline.

3.2.2 Vetores de Contexto

Devido a dificuldade em se obter corpora paralelos, o estudo de como se extrair dicionários bilíngues a partir de corpus comparáveis teve início com o trabalho de Fung (1998), o primeiro a propor a utilização dos contextos das palavras para a sua comparação. Nesse trabalho, são citados 5 problemas pelos quais é mais difícil utilizar corpora comparáveis ao invés de paralelos para essa tarefa:

1. Palavras tem múltiplos significados por corpus.
2. Palavras tem múltiplas traduções por corpus.
3. Traduções podem não existir no documento alvo.
4. Frequências de ocorrência não são comparáveis.
5. Posições não são comparáveis.

Apesar destes fatores, os autores afirmam que, dentro de um mesmo tópico, as palavras tem *contextos* comparáveis, e que para um mesmo domínio e um mesmo período de tempo, tem padrões de uso comparáveis. A partir dessa premissa são realizados experimentos em corpus de jornais em Chinês e Inglês e são obtidos resultados que variam de 30% a 76% de precisão quando o primeiro (top-1) ou os 20 primeiros (top-20) resultados são considerados. Além disso os top-20 resultados são, em sua maioria, palavras relacionadas à tradução correta.

A partir do trabalho de Fung (1998), muitos outros exploraram a metodologia de utilizar o contexto das palavras para caracterizá-las, principalmente para a extração de terminologias multilíngues. A estrutura de dados que guarda o contexto das palavras é denominada em muitos trabalhos como **vetor de contexto**.

Além de serem usados para Extração de Terminologias, os vetores de contexto são muito utilizados na construção de thesaurus, que são listas de palavras agrupadas de acordo com a similaridade de significado, geralmente sinônimos (LIN, 1998).

Um vetor de contexto guarda as palavras que co-ocorrem com a palavra à qual o vetor se refere. A co-ocorrência geralmente pode ser definida de duas maneiras: a primeira, mais simples é o modelo *Bag-of-Words* em que definimos uma janela de n palavras ou frases ao redor da palavra de interesse e todas as palavras dentro dessa janela fazem parte do contexto. A segunda maneira é utilizar relações de dependência gramatical para definir o contexto, por exemplo, relações de sujeito e objeto, ou adjetivos e advérbios modificadores. Dessa forma o contexto não será formado apenas por palavras, mas sim tuplas de palavra-relação, em que palavras iguais, com relações diferentes são elementos distintos no vetor. Para essa forma de construção do contexto é necessário ter um parser capaz de extrair as relações em que estamos interessados, o que não é um recurso disponível para todas as línguas, e também a performance varia bastante de acordo com a língua. Existem muitos trabalhos que estudam a influência da maneira de construção dos vetores de contexto no thesaurus como o de Heylen et al (2008).

Além de definir como o contexto será extraído, é preciso definir uma medida de associação entre cada palavra do contexto e a palavra à qual o vetor se refere, essa medida normalmente é uma medida de associação comum como o PMI, ou Log-Likelihood. Por fim precisamos comparar os vetores de contexto construídos para saber o quão similar eles são entre si, para isso também existem diversas medidas possíveis, como a Similaridade de Cossenos e a Similaridade de Jaccard. Tanto as medidas de associação entre a palavra e as palavras do contexto e as medidas de similaridade também foram estudadas quanto a sua influência na construção de thesaurus (CURRAN; MOENS, 2002).

Neste trabalho estamos interessados em extrair o contexto de termos multipalavra, dessa maneira consideramos a forma mais adequada de construir esse contexto é através do modelo *Bag-of-Words*, pois a grande maioria dos parsers não consegue reconhecer adequadamente termos multipalavra para que as relações gramaticais sejam extraídas corretamente. Quanto as outras medidas de associação e de similaridade, todas elas são facilmente aplicadas a termos multipalavra.

3.2.3 Extração de Terminologia Bilíngue

A extração de terminologia bilíngue foi estudada em diversos trabalhos, como descrito por Déjean, Gaussier e Sadat (2002), ela é baseada na premissa que de se duas palavras são traduções mútuas, as palavras que ocorrem junto com elas, as palavras de contexto, também serão traduções. Baseado nessa premissa, a abordagem básica consiste em construir **vetores de contexto** para as palavras de origem e destino para guardar as palavras de contexto mais importantes. O vetor de contexto da palavra de origem é traduzido utilizando um dicionário bilíngue geral e comparado com o vetor de contexto da palavra de destino. Essa abordagem é baseada na maneira que as similaridades entre termos são construídas em Recuperação de Informação através da distância de cossenos entre os vetores de termos extraídos da matriz de termos-documentos.

O uso de um dicionário bilíngue geral é justificado pelo fato de que se os vetores de contexto forem suficientemente grandes, então alguns de seus elementos deverão pertencer a língua comum e ao dicionário bilíngue. Assim, podemos esperar que o vetor de contexto traduzido da palavra t, seja em média, mais próximo do vetor de contexto da tradução s de t. É importante notar que essa abordagem faz sentido mesmo quando t está presente no dicionário, porque o corpus pode apresentar um uso particular, técnico de t.

A implementação desse método segue os seguintes passos:

1. Para cada palavra w, construir um vetor de contexto utilizando uma janela de algumas frases ao redor de w. Cada palavra i no vetor de contexto de w é pesada com uma medida de associação com w. Nesse trabalho foi escolhida a medida *Log Likelihood Ratio*.
2. Os vetores de contexto das palavras de origem são traduzidos utilizando o dicionário bilíngue, não modificando os pesos de cada palavra. Quando muitas palavras são sugeridas como tradução pelo dicionário, todas elas são utilizadas no vetor de contexto, com o mesmo peso.
3. A similaridade de cada palavra de destino s com relação a cada palavra de origem é computada utilizando similaridade de cossenos.
4. As similaridades são normalizadas para gerar uma tabela de probabilidades de tradução, $P(s|t)$, a probabilidade da palavra s ser a tradução de t.

3.2.4 Extração de Terminologia Inglês-Francês de corpus comparável

O artigo de Daille e Morin (2005) aborda a extração de terminologias a partir de um corpus comparável, visto que existem diversas abordagens que utilizam corpus paralelos, porém é muito difícil obter esse tipo de corpus. A principal abordagem utilizada é a de vetores de contexto, onde cada elemento do vetor representa uma palavra que ocorre dentro de uma janela em volta da palavra a ser traduzida. A tradução é obtida comparando o vetor de origem com cada

vetor de tradução candidato depois que o vetor de origem for traduzido usando um dicionário geral. Existem diversos trabalhos que utilizam o contexto para extrair dicionários bilíngues a partir de corpus comparáveis. Esses trabalhos conseguem extrair traduções com uma precisão que varia de 50% a 91%, porém eles focam em termos simples, aqui o foco é encontrar a tradução de termos multipalavra.

Os desafios de identificar a tradução de termos multipalavra são:

1. Termos simples e termos multipalavra não são sempre traduzidos para um termo do mesmo tamanho. Por exemplo, o termo multipalavra francês *peuplement forestier* é traduzido em inglês para o termo simples *crop*, e o termo francês *essence d'ombre* é traduzido para *shade tolerant species*. Esse problema é bem conhecido como **"fertilidade"** e raramente levado em conta em uma extração de terminologia bilíngue.
2. Quando um termo multipalavra é traduzido em outro termo do mesmo tamanho, o termo alvo não é tipicamente composto da tradução das suas partes. Por exemplo, o termo francês *plantation énergétique* é traduzido em inglês como *fuel plantation* onde *fuel* não é a tradução de *énergétique*. Essa propriedade é conhecida como **"não-composicionalidade"**.
3. Um termo multipalavra aparece em diferentes formas, refletindo variações sintáticas, morfológicas ou semânticas, essas variações devem ser levadas em conta na tradução. Por exemplo, os termos em francês *aménagement de la forêt* e *aménagement forestier* se referem ao mesmo termo e ambos são traduzidos para o inglês como *forest management*.

O processo de extração proposto nesse trabalho busca atacar esses três problemas usando métodos linguísticos e estatísticos. Primeiro os termos multipalavra são identificados na língua de destino e na língua de origem usando um programa de extração de termos monolíngue. Depois o algoritmo de alinhamento estatístico é usado para linkar os termos multipalavra da língua de origem e destino. O algorimto de alinhamento extrai palavras e contextos dos termos e propõe traduções comparando-os.

Para a identificação dos termos multipalavra é usado o ACABIT, um programa de código aberto disponível para Inglês e Francês. Esse programa leva em conta diversas variações dos termos: gráficas, inflexões, sintáticas e morfosintáticas. O ACABIT é aplicado em um corpus após sua tokenização, segmentação em frases e rotulação de PoS e lemas. O ACABIT agrupa todas as variações de um termo e no final retorna não uma lista de candidatos, mas uma lista de conjuntos de candidatos, cada conjunto contendo as variações de um candidato a termo.

Para o alinhamento dos termos, queremos alinhar os termos multipalavra da língua origem com palavras, termos simples ou termos compostos da língua destino. A partir daqui, palavras, termos simples e termos multipalavra são chamados de unidades léxicas.

O alinhamento é feito em quatro etapas. Primeiro são construídos os vetores de contexto para cada unidade léxica, para cada unidade i são contadas as frequências de ocorrência de cada unidade léxica j em uma janela de n frases ao redor de i. Para cada unidade léxica i na língua de origem e na língua de destino, obtemos um vetor de contexto, que contém o conjunto

das unidades que co-ocorrem com i e o número de co-ocorrências. Os vetores de contexto são normalizados usando uma medida de associação como *Mutual Information* ou *Log-Likelihood*. Para reduzir o tamanho dos vetores de contexto, somente as co-ocorrências com os valores de associação mais altos são mantidos.

A segunda etapa é construir os vetores de similaridade. Para cada unidade léxica k a ser traduzida identificamos as unidades léxicas cujos vetores são similares ao vetor de k, v_k. Essa similaridade é calculada com uma das medidas de distância de vetores: distância de Cossenos ou de Jaccard. Assim chamamos de vetor de similaridade da unidade k todas as unidades léxicas cujos vetores de contexto são similares a v_k. Para cada unidade l do vetor de similaridade de k é calculado a medida de similaridade entre v_l e v_k e somente os mais similares são mantidos. Até esse ponto somente os vetores de similaridade da língua de origem foram construídos.

A terceira etapa é a tradução dos vetores de similaridade. Usando um dicionário bilíngue, as unidades léxicas dos vetores de similaridade são traduzidas e os vetores de contexto na língua alvo são identificados. Caso a unidade léxica seja um termo simples são gerados tantos vetores de contexto na língua alvo quantas traduções existem no dicionário, então é calculada a união desses vetores para obter um único vetor de contexto alvo. Se a unidade léxica for um termo multipalavra, são gerados tantos vetores de contexto quantas combinações de traduções das partes forem identificadas pelo ACABIT e calculada a união dos vetores.

A quarta e última etapa é encontrar as traduções dos termos multipalavra, para isso é calculado o centro de todos os vetores de contexto na língua alvo, chamado de vetor médio alvo. A tradução de uma unidade léxica é a unidade léxica mais próxima do vetor médio alvo de acordo com a distância de vetores.

O corpus utilizado nesse trabalho tem 4 milhões de palavras e trata de florestas e indústrias florestais. O dicionário bilíngue utilizado tem 22.300 palavras de linguagem geral, com uma média de 1,6 traduções por palavra.

Para avaliação dessa metodologia foram utilizados glossários e thesaurus de diferentes fontes, deles foram extraídas 3 listas de termo para avaliação:

1. 100 termos simples em francês cuja tradução é um termo simples em inglês. A tradução desses termos não é dada pelo dicionário bilíngue.
2. 100 termos multipalavra em francês cuja tradução pode ser um termo simples ou multipalavra em inglês. No caso de termos multipalavra, a tradução do termo não pode ser obtida pela tradução das partes.
3. 100 termos multipalavra em francês cuja tradução é dada pela tradução das partes.

Os parâmetros do método com os quais foram obtidos os melhores resultados foram os seguintes: uma janela de contexto de 3 frases ao redor da unidade léxica a ser traduzida, os vetores de contexto foram construídos somente com palavras unitárias e foram limitados aos 100 maiores valores de *Log-Likelihood*. Os vetores de similaridade foram contruídos com os 30 menores valores de distância de cossenos e a tradução também é encontrada usando a distância

de cossenos.

Os melhores resultados foram obtidos para a terceira lista de palavras, com 89 das 100 palavras tendo suas traduções encontradas e quase sempre em uma das primeiras posições a posição média foi 3,8. Para os outros dois grupos de palavras os resultados não foram tão bons, com 56 e 63 das 100 traduções encontradas para as listas 1 e 2, com uma posição média de 32,9 e 30,7. Os autores comentam que especialmente para as listas de palavras 1 e 2 diversos resultados diferentes foram encontrados para diferentes configurações dos parâmetros com traduções para certas palavras sendo encontradas somente em uma configuração e para outras com outra configuração.

3.2.5 Revisando o método composicional para aquisição de terminologia a partir de corpora comparáveis

O artigo de Daille e Morin (2012) faz uma comparação entre dois métodos de extração de terminologia a partir de corpora comparáveis. O mais simples deles utiliza um dicionário bilíngue como base para gerar as traduções candidatas e o método proposto no artigo que utiliza o contexto das palavras que não se encontram no dicionário.

Conforme apontado por outros trabalhos (NAKAGAWA; MORI, 2003) e (SAVARY; JACQUEMIN, 2003) os termos multipalavras de um domínio específico são mais representativos e menos polissêmicos[1] que termos simples e representam cerca de 80% dos termos de domínio específico, por isso a importância deles para a criação de uma terminologia.

A abordagem composicional é o método mais simples para aquisição de terminologia, baseia-se na propriedade que "O significado do todo é uma função do significado das partes"[2] (KEENAN; FALTZ, 1985) e é composta por três etapas:

1. **Tradução da multipalavra de origem** Cada uma das palavras que compõem as multipalavras da língua original são traduzidas por um dicionário bilíngue. Todas as traduções possíveis são armazenadas.
2. **Geração das traduções candidatas** São construídas todas as possíveis combinações de termos na língua destino usando as traduções geradas pelo dicionário, não importando a ordem das palavras. Dessa forma são geradas $\Theta((\prod_{i=1}^{n} t_i)n!)$ possíveis candidatos a tradução, onde t_i é o número de traduções da palavra i do termo e n o número de palavras. Esse número pode ser reduzido usando padrões de PoS, por exemplo, se temos um termo na língua de origem no padrão N N, sabemos que a tradução desse termo vai ter um padrão N N ou A N.
3. **Seleção das traduções candidatas** As traduções mais prováveis são selecionadas de acordo com a frequência na língua destino.

[1] Polissemia é o fato de uma determinada palavra ou expressão ter mais de um significado, por exemplo a palavra jaguar pode se referir ao animal ou à marca de carros.

[2] Do original, "the meaning of the whole is a function of the meaning of the parts"

Esse algoritmo é simples mas falha em algumas situações: (i) um dos elementos do termo multipalavra não é encontrado no dicionário bilíngue; (ii) a combinação traduzida é válida mas não foi encontrada pelo extrator de termos na língua destino, o que pode acontecer caso o termo não ocorra no corpus de destino, ou então algum erro no pré-processamento tenha excluído o termo dos resultados. (iii) a combinação traduzida não é válida devido a um dos problemas descritos anteriormente em 3.2.4: fertilidade, não-composicionalidade e variação dos termos.

O método proposto neste artigo, chamado Método Composicional com Projeção Baseada em Contexto, é capaz de resolver os pontos (i) e (iii), e é composto de quatro etapas:

1. **Extração do Contexto do Termo multipalavra** Um termo multipalavra é definido como $C_{s1}C_{s2}...C_{sk}$ onde cada C_{si} é uma palavra que compõe o termo, e k é o número de palavras no termo. Buscamos no dicionário bilíngue por cada C_{si}, caso não seja encontrada substituímos ela por informação de co-ocorrência. A informação de co-ocorrência é computada entre a palavra C_{si} e as palavras que co-ocorrem em uma janela de w palavras ao redor dela. As medidas usadas para guardar a informação de co-ocorrência são *Mutual Information* ou *Log-likelihood*. Como exemplo na figura 3.1 (MORIN; DAILLE, 2012) temos o termo multipalavra francês *antécédent familial* onde a primeira palavra C_{s1} não existe no dicionário bilíngue e é substituída pelo seu vetor de contexto V_{s1}.

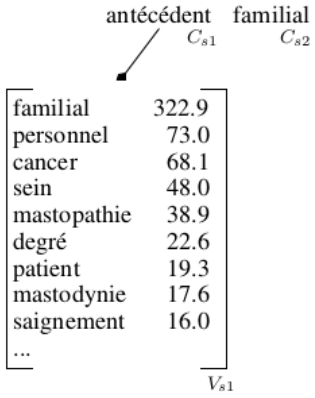

Figura 3.1 – Vetor de contexto para *antécédent familial*

2. **Transferência dos termos multipalavra para a língua destino** Nessa etapa podem ocorrer duas situações dependendo se a tradução da palavra componente do termo foi encontrada no dicionário ou não:
   i) Se o componente C_{si} foi encontrado, computamos os vetores de contexto (V'_{si}) de cada tradução no corpus de destino.
   ii) Se o componente C_{si} não estava no dicionário, usamos o vetor de contexto extraído do corpus de origem na etapa anterior. Os elementos do vetor V_{si} são projetados no corpus destino usando o dicionário bilíngue e se torna V'_{si}. Caso exista mais de uma tradução para um elemento, todas elas são usadas, mas são pesadas de acordo com sua frequência no corpus de destino. Caso um elemento não seja encontrado no dicionário, ele é descartado.

Na figura 3.2 (MORIN; DAILLE, 2012) estão ilustrados as duas situações descritas, (i) para *familial* e (ii) para *antécédent*

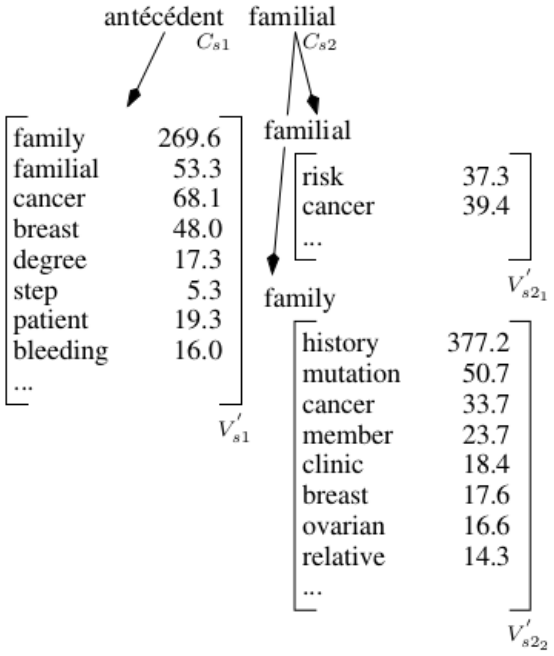

Figura 3.2 – Vetores de contexto na língua destino para *antécédent familial*

3. **Geração das Traduções Candidatas** Cada termo multipalavra da língua destino, da qual cada componente C_{ti} é descrito por seu vetor de contexto V_{ti} é comparado aos termos multipalavra transferidos para a língua destino na etapa anterior usando uma medida de similaridade como Jaccard ou similaridade de Cossenos. São calculadas todas as combinações de similaridade possíveis, por exemplo, se estamos comparando dois termos compostos de duas palavras com seus vetores na língua destino V_{t1} V_{t2} e os vetores transferidos V'_{s1} e V'_{s2}, as possíveis combinações são $sim(V'_{s1}, V_{t1})$ e $sim(V'_{s2}, V_{t2})$ ou $sim(V'_{s1}, V_{t2})$ e $sim(V'_{s2}, V_{t1})$. O score final é dado pela média geométrica das similaridades, no exemplo, os dois scores seriam: $\sqrt{sim(V'_{s1}, V_{t1})sim(V'_{s2}, V_{t2})}$ e $\sqrt{sim(V'_{s1}, V_{t2})sim(V'_{s1}, V_{t2})}$. Caso exista mais de uma tradução possível para determinada palavra, é criada uma combinação para cada possível tradução. Na figura 3.3 (MORIN; DAILLE, 2012) são ilustradas as comparações feitas entre os vetores de *antécédent familial* e *family history*.
4. **Ordenamento das traduções candidatas** As traduções candidatas são ordenadas de acordo com o score de similaridade.

Os corpora utilizados eram do domínio médico com sub-domínio "câncer de mama" (*breast cancer*). Os documentos foram coletados automaticamente de sites com artigos científicos. No total foram coletados 118 documentos para inglês, 130 para francês e 103 para para alemão, contendo 530.000 palavras para Inglês e Francês e 220.000 para alemão.

No trabalho de Daille e Morin foram feitos experimentos com o alinhamento de termos

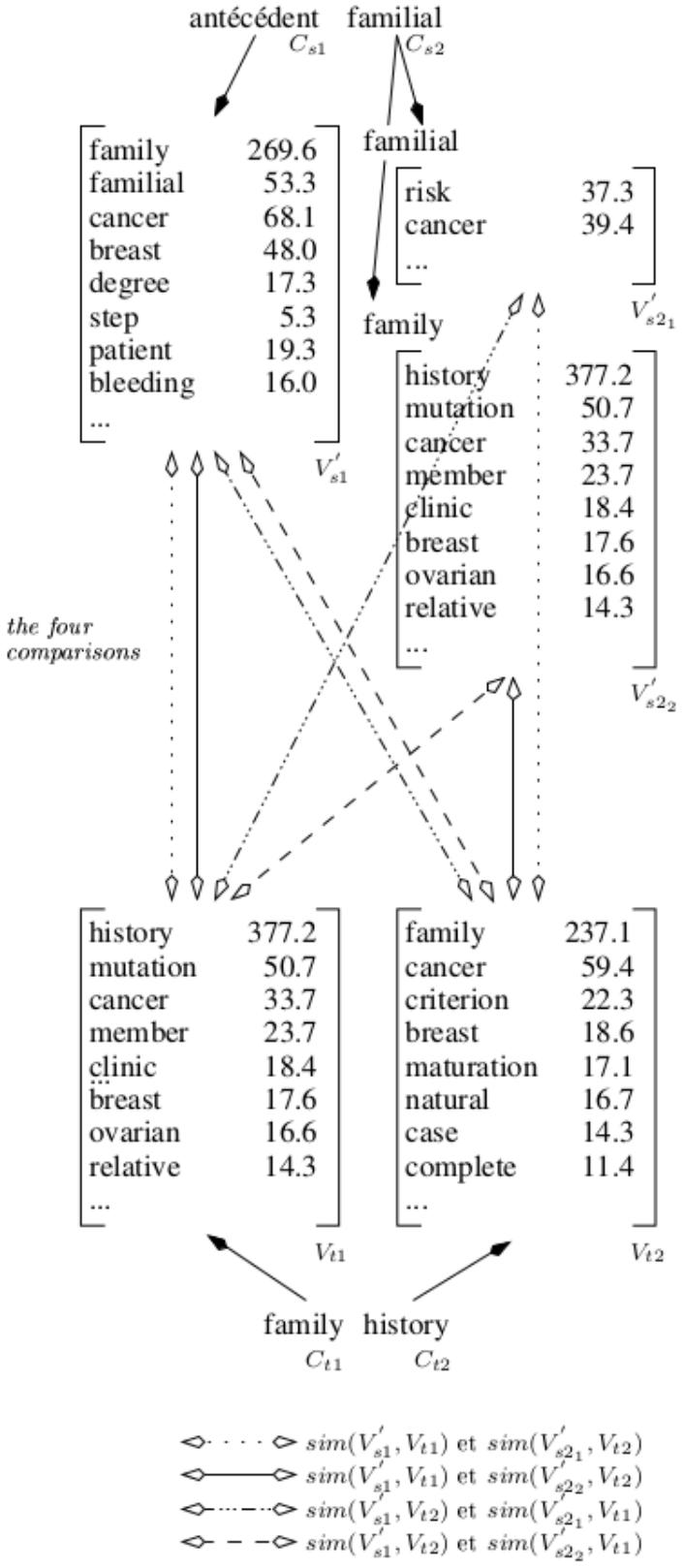

Figura 3.3 – Comparações feitas entre os vetores de *antécédent familial* e *family history*

para francês/inglês e francês/alemão, foram utilizados dicionários bilíngues contendo 243.539 e 170.967 traduções respectivamente para cada par de línguas.

Para a extração monolíngue de termos foi utilizado o TTC TermSuite (ROCHETEAU; DAILLE, 2011), que aplica o mesmo método de extração de termos para diversas línguas.

O conjunto de testes foi construído com os termos multipalavra em Francês extraídos pela ferramenta que tinham frequência de ocorrência no corpus maior ou igual a 5. Esse conjunto é composto de 976 termos, dos quais 90% são formados por somente duas palavras.

Os experimentos realizados foram divididos nas seguintes etapas:

Busca no dicionário: Foram contados o número de termos traduzidos diretamente buscando no dicionário. Dos 976 termos, 51 estavam no dicionário Francês/Inglês e 12 no Alemão/Inglês. Esses termos são em sua maioria genéricos, como *medical treatment* e *amino acid*.

Método Composicional: Como esse método pode gerar mais de uma tradução candidata foi calculada a precisão para o primeiro (Top_1) e para os cinco primeiros resultados (Top_5) recuperados. Com esse método foram geradas 140 traduções dos termos em Francês, com uma precisão de 73.2% para o Top_1 79.1% para o Top_5. Para os termos em alemão foram geradas 87 traduções com uma precisão de 88.8% para o Top_1 e 95.7% para o Top_5.

Método Composicional com Projeção Baseada em Contexto: As melhores configurações encontradas para esse método foram uma janela de contexto de 3 palavras, *Mutual Information* como medida de associação e Similaridade de Cossenos. Para esse método foram calculados os valores de precisão para os Top 1, 5, 10 e 20 resultados. Foram encontradas traduções para 514 termos em Francês com uma precisão entre 42.1% (Top_1) e 57.1% (Top_{20}), para o Alemão foram encontradas traduções para 510 termos com uma precisão entre 44.3% (Top_1) e 51.2% (Top_{20}).

3.2.6 Melhorando a Extração de Léxicos Bilíngues a partir de corpora comparáveis usando modelos baseados em janela e sintaxe

O trabalho de Hazem e Morin (2014) procura comparar dois modelos de representação dos vetores de contexto de uma palavra na tarefa de identificação de traduções em um corpus comparável. Eles comparam o modelo baseado em janela, que utiliza um número fixo de palavras ao redor da palavra alvo para representação desse contexto, e o modelo baseado em sintaxe que utiliza relações de dependência extraídas utilizando regras de sintaxe (geralmente utilizando um parser). Hazem e Morin (2014) também propõem uma maneira de combinar os dois modelos (baseado em janela e sintaxe) para se obter resultados melhores.

O modelo baseado em janela funciona de maneira simples, são contadas todas as palavras que ocorrem ao redor da palavra alvo para todas as ocorrências dela no corpus, assim formando

o vetor de contexto dela. O modelo baseado em sintaxe usado aqui, é baseado no trabalho de Otero (2008), que lista 7 relações de dependência entre palavras. Nesse modelo a informação do contexto não consiste em apenas palavras, mas de tuplas formadas por uma palavra e a relação dela com a palavra alvo, assim em um vetor de contexto teremos o número de vezes que uma palavra aparece como sujeito (ou outra relação de dependência) da palavra alvo.

A abordagem de combinar os dois modelos é proposta pois no modelo baseado em janela todas as palavras tem a mesma importância, o que é inconsistente e nos leva a procurar pelo modelo baseado em sintaxe, porém esse modelo necessita de grandes corpora para ser eficiente, o que não é sempre verdade quando trabalhamos com corpora comparáveis de domínio específico. Assim duas abordagens de combinação dos dois modelos foram propostas:

- Pós-combinação: Baseado nas técnicas de Recuperação de Informação que combinam os resultados de diferentes engines para melhorar a performance, a pós-combinação utiliza os rankings de similaridade produzidos pelos dois modelos e cria um novo utilizando uma combinação aritmética simples dos scores de similaridade.
- Pré-combinação: Como as duas representações do contexto não são exclusivas, a pré-combinação usa informações do modelo baseado em janela, com a contagem das palavras que co-ocorrem com a palavra alvo w, e infomações do modelo baseado em sintaxe, com a contagem das palavras que tem uma relação de dependência com w. Assim uma palavra que não tem relações de dependência com w podem estar no contexto baseado em janela, ajudando a ter uma cobertura maior. Essa abordagem se baseia na hipótese que os modelos são complementares.

Como recursos linguísticos foram usados quatro corpora comparáveis Inglês-Francês contendo entre 200.000 e 500.000 palavras, aproximadamente 5.000-9.000 distintas (para cada língua). Para avaliação foram selecionados termos simples que ocorrem mais de 4 vezes no corpus, para cada corpus foram selecionados entre 150 e 321 termos. O parser utilizado foi disponibilizado por Otero (2008).

Os testes realizados utilizaram combinações entre as medidas de associação entre as palavras de contexto e a palavra alvo (PMI, Log-Likelihood) e as medidas de similaridade entre os vetores (similaridade de cossenos, distância de Jaccard). Os resultados foram avaliados utilizando MAP (*Mean Average Precision*, equação 6.1) e estão descritos na tabela 3.1, onde cada coluna representa um dos quatro corpora utilizados (sobre diabetes, vulcões, vento-energia e câncer de mama).

Os resultados mostraram que o modelo baseado em janela foi melhor que o modelo baseado em sintaxe em todas as configurações testadas, o que se deve a falta de dados provenientes de um corpus de domínio e a erros de *parsing*. Para todas as configurações, exceto a que utiliza PMI como medida de associação, a pré-combinação melhorou significativamente os resultados obtidos. Os resultados utilizando PMI não foram melhores pois o PMI tende a superestimar baixas frequências e como o modelo baseado em sintaxe introduz muitas relações de dependên-

Tabela 3.1 – Resultados usando janela, sintaxe e combinações das duas abordagens

		Diabetes	Vulcões	Vento-energia	Câncer de Mama
PMI-Cosseno	janela	15,5	22,5	15,7	22,4
	sintaxe	10,4	22,0	14,1	15,7
	Pré-comb.	5,5	17,3	12,2	15,0
	Pós-comb.	26,6	42,3	26,8	29,2
DOR-Cosseno	janela	16,5	37,5	21,3	24,4
	sintaxe	10,0	14,9	10,6	12,0
	Pré-comb.	23,7	45,2	**34,9**	34,7
	Pós-comb.	23,6	44,0	31,8	32,1
LL-Jaccard	janela	20,6	49,7	29,0	29,3
	sintaxe	17,8	28,5	21,9	18,0
	Pré-comb.	27,0	51,2	32,5	**36,1**
	Pós-comb.	**27,8**	**53,8**	33,3	33,9

cia com baixas frequências, elas foram superestimadas. A pós-combinação também foi melhor em todas as configurações, incluindo a que utilizou PMI como medida de associação. Os resultados obtidos apóiam a hipótese que os dois modelos são complementares e mais adequados a pequenos corpora.

4 MATERIAIS E MÉTODOS

Este capítulo descreve os recursos e ferramentas utilizadas neste trabalho, na seção 4.1 descrevemos os corpora utilizados para a avaliação dos resultados obtidos, na seção 4.2 descrevemos as ferramentas utilizadas neste trabalho.

4.1 Corpora

Os corpora utilizados nesse trabalho foram o GENIA(OHTA; TATEISI; KIM, 2002) e o Cameleon(GRANADA et al., 2012).

4.1.1 GENIA

O GENIA[1] é um corpus em Inglês sobre biologia molecular, contendo 1999 resumos do MEDLINE (*Medical Literature Analysis and Retrieval System Online*, ou Sistema Online de Busca e Análise de Literatura Médica), a base de dados bibliográficos da Biblioteca Nacional de Medicina dos Estados Unidos. Os textos do GENIA foram selecionados utilizando o *PubMed*[2], o serviço que permite acesso aos textos da Medline, usando uma consulta por *"human", "blood cells"* e *"transcription factors"* .

O GENIA contém diversas anotações sintáticas e semânticas no corpus, das quais as importantes para este trabalho são o corpus anotado com PoS e o corpus anotado com os termos. O PoS é utilizado para facilitar o processamento, (eliminando a etapa de anotação com o PoS do pré-processamento) e os termos anotados são utilizados como gold-standard para avalição dos resultados.

O GENIA contém 490.752 tokens e é pequeno em comparação com outros corpora como o Europarl (40 milhões de tokens) e o BNC (100 milhões). Porém o fator mais importante do GENIA são os termos anotados, são 97.876 termos anotados, dos quais 55.487 são termos multipalavra (esses números incluem diversas instâncias de cada termo). Os termos encontrados no corpus podem ser arbitrariamente grandes, com o maior termo tendo 23 tokens. Na análise feita por Ramisch (2009) cada frase do GENIA tem, em média, 5 termos, dos quais 3 são multipalavra. Isso demonstra o quão específica é a informação presente nesses textos, e a importância de se tratar corretamente os termos multipalavra em um domínio específico.

4.1.2 Cameleon

O corpus do projeto Cameleon[3] foi coletado automaticamente na web buscando por páginas com o domínio de conferências, utilizando como base uma lista de termos multipalavra

[1] Disponível em <http://www.nactem.ac.uk/genia/home>
[2] Disponível em <http://www.ncbi.nlm.nih.gov/pubmed/>
[3] Disponível em: <http://cameleon.imag.fr/>

sementes que foram manualmente selecionados. A metodologia de construção do corpus está descrita no trabalho de Granada et al (2012). O corpus do Cameleon é um corpus comparável que contém textos em Inglês, Português e Francês. Na tabela 4.1 temos os dados das línguas utilizadas nesse trabalho.

Tabela 4.1 – Corpus Cameleon

Língua	Tokens	Tipos
Português	16.763.582	372.940
Inglês	16.966.728	331.299

4.2 Ferramentas

A principal ferramenta utilizada como auxílio foi o *Ngram Statistics Package (Text-NSP)* (BANERJEE; PEDERSEN, 2003). O Text-NSP[4] é um software de código livre, escrito em Perl que disponibiliza o cálculo de diversas métricas de uma maneira simples.

Também utilizamos o TreeTagger (SCHMID, 1994) para o pré-processamento do texto e extração do *part-of-speech* (PoS), utilizado para a filtragem inicial dos candidatos a termos.

4.2.1 Text-NSP

O Text-NSP possui dois scripts principais, o count.pl e o statistic.pl. O primeiro deles, count.pl serve para calcular as frequências dos n-gramas encontrados em um arquivo texto. A maneira mais simples de utilizá-lo é a seguinte:

```
count.pl output.cnt input.txt
```

Onde o arquivo input.txt contém o texto de entrada do qual se quer contar as frequências dos n-gramas e output.cnt será o arquivo que conterá a saída. O programa count.pl faz a tokenização do arquivo de entrada e depois agrupa os tokens em n-gramas. O tamanho default dos n-gramas computados é 2, mas é possível modificar esse parâmetro e outros na chamada do programa. Alguns dos parâmetros disponíveis são:

- -ngram N Cria n-gramas de N tokens cada.

- -token FILE Usa expressões regulares em FILE para tokenizar o arquivo.

- -nontoken FILE Remove todos os caracteres especificados por meio de expressões regulares em FILE.

- -stop FILE Remove da contagem os n-gramas contendo as expressões regulares especificadas em FILE.

[4]Disponível em <http://search.cpan.org/dist/Text-NSP/>

Os parâmetros acima foram utilizados no desenvolvimento desse trabalho. O tamanho dos n-gramas extraídos foi utilizado para fazer a contagem de bigramas e trigramas. Um arquivo de tokenização específico foi utilizado para cada um dos corpus utilizado, pois cada tipo de texto possui tokens com características próprias, um texto de domínio geral é geralmente tokenizado considerando caracteres alfanuméricos, porém textos de domínio específico podem possuir caracteres especiais em seus tokens como, por exemplo, os termos do GENIA "ca2+-modulating cyclophilin ligand" e "v-(d)-j recombinase activity" que possuem caracteres +-() que não são geralmente considerados parte de um token. O parâmetro nontoken foi utilizado para remover caracteres como pontuações dos tokens e por fim o arquivo stop foi utilizado para remover os n-gramas que contivessem palavras que conhecidamente não fazem parte de termos, como artigos e conjunções muito comuns.

O segundo programa do Text-NSP utilizado é o statistic.pl, ele é usado para calcular medidas de associação entre os n-gramas. A maneira default de utilizá-lo é a seguinte:

```
statistic.pl medida output.txt input.cnt
```

Onde o input.cnt é o arquivo de saída com a contagem de frequências dos n-gramas produzida pelo count.pl, output.txt é o arquivo de saída produzido pelo statistic.pl e medida é a medida de associação que se deseja calcular, dentre as diversas disponíveis no Text-NSP[5], sendo as mais comuns o coeficiente de Dice, o coeficiente de Jaccard, o PMI e o *Log-Likelihood Ratio*.

4.2.2 TreeTagger

O TreeTagger[6] é uma ferramenta para anotar o texto com *part-of-speech* e *lema*, ele pode ser usado em diversos idiomas como inglês, português, francês, entre outros.

Lema é a forma canônica de uma palavra, ou seja, a forma não flexionada dela. A forma canônica de um substantivo é o seu singular, por exemplo, a lematização de "participantes" será "participante". A forma canônica de um verbo será seu infinitivo, por exemplo, a lematização de "correm" será "correr".

Outro aspecto importante do TreeTagger é a sua velocidade ao tratar grandes conjuntos de textos quando comparado a outros taggers (em uma avaliação para o Inglês o TreeTagger foi 2,9 vezes mais rápido que o segundo colocado, processando mais de 20000 palavras por segundo)[7]. O TreeTagger também funciona através da linha de comando e requer apenas o arquivo de entrada como parâmetro.

```
tree-tagger-english input
```

[5] As medidas de associação disponíveis no Text-NSP estão descritas em <http://search.cpan.org/~tpederse/Text-NSP/doc/README.pod>

[6] Disponível em <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

[7] Esta avaliação detalhada da performance de diversos *taggers* pode ser encontrada em <http://mattwilkens.com/2008/11/08/evaluating-pos-taggers-speed/>

A saída é escrita na terminal e tem o seguinte formato: três colunas, sendo a primeira a palavra original, a segunda o PoS e a terceira o lema.

```
The         DT    the
TreeTagger  NP    TreeTagger
is          VBZ   be
easy        JJ    easy
to          TO    to
use         VB    use
```

4.2.3 Lucene

O Lucene[8] é uma biblioteca de indexação e pesquisa de textos escrita em Java desenvolvida pela Apache. O Lucene utiliza uma mistura do modelo vetorial e do modelo booleano para atribuir scores aos documentos em suas buscas. Ele usa o modelo booleano antes, para diminuir o número de documentos aos quais será preciso atribuir um score, mas essencialmente utiliza o modelo vetorial.

O modelo booleano de Recuperação de Informação é um modo de indexação de documentos no qual são armazenadas os documentos em que cada palavra aparece, assim o resultado de uma pesquisa é uma lista de documentos que contém a determinada palavra. No modelo booleano é possível utilizar os operadores booleano (*and, or, not*) para fazer consultas mais complexas. O modelo vetorial utiliza o tf-idf para, a partir de uma consulta, retornar uma lista de documentos ordenada de acordo com a importância deles, calculada pelo tf-idf, onde o TF (*term frequency*) dá importância para termos que aparecem muito, e o IDF (*inverse document frequency*) dá importância aos termos que ocorrem em poucos documentos, ou seja o tf-idf dá mais importância a termos que ocorram muitas vezes em poucos documentos.

O Lucene possui vários recursos de pesquisa, como:

- Pesquisa ranqueada (os melhores resultados retornados primeiro);
- Consultas de frase, consultas wildcard, consultas de proximidade, consultas de range;
- Pesquisa por campos (por ex., título, autor, conteúdo);
- Pesquisa por intervalo de datas;
- Ordenação por qualquer campo;
- Pesquisa de múltiplos índices com merge de resultados;
- Permite simultaneamente fazer atualização e pesquisa.

O Lucene possui diversos tipos de analisadores, que são classes que especificam como será o pré-processamento do texto. Alguns desses analisadores são:

[8] Disponível em <http://lucene.apache.org/core/>

- WhitespaceAnalyzer: um analisador muito simples que apenas tokeniza usando os espaços em branco;
- StopAnalyzer: remove as *stopwords*;
- SnowballAnalyzer: faz *stemming*. A técnica de *Stemming* consiste em reduzir as palavras aos seus radicais, permitindo que palavras morfologicamente relacionadas sejam representadas por uma única forma comum. Por exemplo, as palavras "aprendeu", "aprendendo" e "aprendo" podem ser reduzidas ao radical "aprend".;
- StandardAnalyzer: faz filtragem de stopwords e lower case, e ainda tenta fazer uma limpeza nas palavras, por exemplo, retirando apóstrofes (') removendo pontos de siglas ("T.L.A." para "TLA");
- SimpleAnalyser: Divide o texto em caracteres que não são letras e coloca em *lowercase*.

Neste trabalho o Lucene é utilizado para encontrar todas as ocorrências de uma EM em um corpus, e assim construir o seu vetor de contexto.

5 ARQUITETURA

Este capítulo descreve o sistema desenvolvido para extração e alinhamento de termos. Na seção 5.1 descreveremos o sistema de extração, com suas etapas de pré-processamento e métodos implementados e na seção 5.2 é descrito o sistema para alinhamento dos termos.

5.1 Extração de Termos

5.1.1 Arquitetura Geral

Na figura 5.1 podemos ver as etapas do sistema de extração de termos multipalavra, que serão detalhadas nas próximas seções.

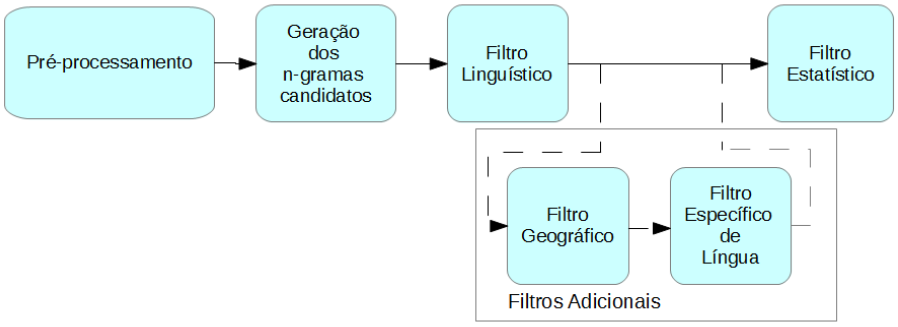

Figura 5.1 – Etapas da Extração

O sistema desenvolvido tem como entrada um corpus, que pode ser um texto em sua forma de superfície ou um corpus anotado como o GENIA. Para que essa flexibilidade fosse possível foi desenvolvida uma classe abstrata responsável pela leitura de um corpus e para cada tipo de corpus disponível foi criada uma classe que cuida dos detalhes específicos de cada corpus. Após a leitura do corpus é feito o pré-processamento necessário àquele tipo de corpus. Por exemplo, para um corpus de texto em sua forma de superfície é necessário utilizar o TreeTagger para extrair o *PoS* das palavras, já para um corpus anotado, como o GENIA, é necessário somente ler do arquivo o *PoS* das palavras.

Após a leitura do corpus e do pré-processamento é feito a contagem de frequência e o cálculo das medidas de associação dos n-gramas utilizando o Text-NSP. De posse de todos os n-gramas presentes no texto, é feita a filtragem por *PoS*, que reduz o número de candidatos a serem analisados, e por fim é feito o cálculo das outras métricas implementadas, o c-value e o nc-value. A saída do sistema apresenta os candidatos a termos multipalavra encontrados e todas as métricas calculadas, assim o usuário pode ordená-los por qualquer uma das métricas para analisar os resultados.

5.1.2 Pré-Processamento

O pré-processamento do texto é composto por duas etapas: a tokenização e a etiquetação morfossintática(*POS Tagging*). A tokenização é a separação do texto em tokens, que para o português e inglês são basicamente palavras, assim separamos a pontuação das palavras no texto. A tokenização pode ser diferente para cada corpus utilizado, em geral consideramos como um token uma sequência de letras, porém para um corpus específico como o GENIA, podemos ter caracteres especiais fazendo parte das palavras, como em *"ca2+-modulating cyclophilin ligand"* e *"v-(d)-j recombinase activity"*.

A segunda etapa, a etiquetação morfossintática (em inglês, *POS Tagging*) inicia com a execução do TreeTagger, que anota o corpus com o PoS e o lema de cada palavra. A partir da saída do TreeTagger são salvas três importantes informações associadas a cada palavra do corpus: frequência, PoS e lema, que serão usadas nas próximas etapas da Extração e também no Alinhamento.

5.1.3 Geração dos N-gramas Candidatos

Após o pré-processamento, o Text-NSP (seção 4.2.1) é usado para gerar todos os possíveis n-gramas[1] presentes no corpus, e também calcular as seguintes medidas de associação: *Log Likelihood Ratio*, *Pointwise Mutual Information* (PMI), *Mutual Information* e *Poisson Stirling Measure* que serão utilizadas para ordenar os termos candidatos.

5.1.4 Filtro Linguístico

Após a extração dos n-gramas feita pelo Text-NSP, é feita a filtragem por PoS usando os padrões definidos por Justeson e Katz (JUSTESON; KATZ, 1995) em termos de substantivos (N), adjetivos (J) e preposições (P) e refinados após testes, onde os padrões que não encontravam termos interessantes foram excluídos. Como neste trabalho estamos interessados em termos nominais, os padrões são combinações de substantivos e adjetivos. Essa filtragem verifica em qual padrão se encaixa cada n-grama e exclui aqueles que não se encontram nos padrões aceitos.

Os padrões utilizados para o Português e para o Inglês podem ser vistos nas tabelas 5.1 e 5.2.

Essa etapa de filtragem possibilita descartarmos combinações muito comuns de artigos e preposições e também n-gramas envolvendo verbos. De posse dessa lista filtrada de candidatos, podemos calcular as outras métricas.

[1]Escolhemos utilizar somente n=2 e n=3 pois a grande maioria dos termos multipalavra tem esse tamanho, e também por questões de tempo de processamento

Tabela 5.1 – Padrões de PoS - Português

Padrão	Exemplo
N N	data limite, conselho fiscal
N J	comissão organizadora, desenvolvimento sustentável
N N N	Pontifícia Universidade Católica
N N J	internet banda larga, igreja católica romana
N J J	produto interno bruto, congresso latino americano
N P N	tecnologia da informação, ponto de vista

Tabela 5.2 – Padrões de PoS - Inglês

Padrão	Exemplo
J N	international conference, social media
J N N	best paper award, national science foundation
N N	conference call, poster session
N N N	paper submission deadline, program committee member

5.1.5 Filtros Adicionais

Além do filtro linguístico, adicionamos mais dois filtros específicos para este trabalho: o filtro geográfico e um filtro específico de língua. Estes filtros foram incluídos após diversos testes durante a etapa do alinhamento para melhorar a performance, chamamos eles de filtros adicionais pois eles não são necessários para o funcionamento do sistema e, se analisarmos apenas a etapa de extração, a inclusão desses filtros exclui resultados interessantes e provavelmente desejados, como nomes de locais (filtro geográfico) e expressões estrangeiras comuns (filtro linguístico), porém estes termos dificilmente têm correspondentes na outra língua, o que dificulta a tarefa de alinhamento.

5.1.5.1 Filtro Geográfico

Adicionamos um filtro geográfico para excluir nomes de cidades, estados, países e regiões. Para isso armazenamos uma lista com todos os nomes de locais que queremos excluir e caso um dos termos que passou no filtro linguístico esteja nesta lista, ele é excluído. Essa lista de locais foi obtida do site Geonames[2] que disponibiliza gratuitamente para download uma base de dados com aproximadamente 10 milhões de nomes principais e alternativos de locais geográficos. Esses locais estão divididos em nove categorias, das quais estamos interessados nas seguintes: países/estados/regiões e cidades/vilarejos. Outras categorias como montanhas/colinas, lagos/rios e estradas também poderiam ser usadas, porém em testes realizados utilizando todas as categorias possíveis, nomes de termos interessantes foram excluídos, como por exemplo, “access point”, que se refere a um ponto de acesso à internet no domínio de

[2]Disponível em <www.geonames.org>.

informática, também é um nome de um ponto rochoso localizado na Antártida.

A inclusão desse filtro geográfico se deve ao fato de que muitos nomes de cidades e países são identificados utilizando esse método de extração de termos, muitos deles sendo ordenados em altas posições. Entretanto como o objetivo final deste trabalho é identificar termos multi-palavra e seus equivalentes em outra língua em corpora comparáveis, observamos que dificilmente encontramos as traduções de nomes de locais nos corpora da outra língua. Por exemplo, o nome de um país muito comum com Estados Unidos, é relativamente fácil de encontrar o equivalente *United States*, porém quando procuramos por nomes de cidades, os nomes mais frequentes variam muito entre os corpora de línguas diferentes, assim em um corpus em português encontraremos nomes de cidades como São Paulo, Rio de Janeiro, e em um corpus em inglês, *San Francisco, Los Angeles*.

5.1.5.2 Filtro Específico de Língua

Como os corpora comparáveis utilizados nesse trabalho são obtidos automaticamente através de um crawler (um programa que navega pela internet de forma automatizada, nesse caso coletando páginas da web para serem usadas como recurso linguístico), eles contém algum tipo de ruído como, por exemplo, algumas páginas que contém conteúdo em uma língua diferente da qual desejamos. Para contornar esse problema criamos dois filtros para que tenhamos certeza que o termo que estamos analisando é da língua desejada. Para o inglês utilizamos a WordNet[3](uma base léxica das palavras da língua inglesa que provê informações sobre similaridade sintático semânticas) e para o português utilizamos o léxico de Português Brasileiro construído nos trabalhos de Muniz (2003) e (2004).

Esse filtro verifica individualmente cada palavra contida em um termo que passou pelos filtros anteriores, e caso uma das palavras não esteja presente no dicionário, ela é descartada.

5.1.6 Métricas

O último passo é o cálculo de métricas para a ordenação dos termos. Como o foco desse trabalho é o alinhamento, apenas algumas das métricas mais utilizadas nos trabalhos descritos no capítulo 3 foram implementadas. Qualquer uma dessas métricas pode ser usada na ordenação final dos termos candidatos. Os resultados de cada métrica são discutidos na seção 6.1.

5.1.6.1 Frequência e Métricas de Associação

A frequência é o método mais simples, ela é simplesmente a contagem do número de vezes que o termo aparece no corpus. As métricas de associação utilizadas nesse trabalho são *Log Likelihood Ratio, Pointwise Mutual Information, Mutual Information* e *Poisson Stirling Mea-*

[3]Disponível em <http://wordnet.princeton.edu/>

sure Nesse trabalho a frequência e as medidas de associação foram calculadas utilizando o Text-NSP.

5.1.6.2 c-value

O c-value (FRANTZI; ANANIADOU; TSUJII, 1998) foi implementado em Java. Para calculá-lo começamos obtendo as frequências dos termos filtrados por PoS, e após isso, para cada termo verificamos quais os termos que o contém e diminuimos a frequência média desses termos do termo atual, como visto na equação 3.1. Essa medida terá um valor alto para os termos maximais, aqueles que não estão contidos em nenhum outro candidato, por exemplo, o termo "tribunal penal internacional" terá um c-value mais alto que "tribunal penal".

5.1.6.3 nc-value

O nc-value (FRANTZI; ANANIADOU; TSUJII, 1998) é uma extensão do c-value que busca utilizar informação de contexto para melhorar a extração de termos feita pelo c-value. O nc-value se divide em duas etapas: o cálculo do score para as palavras adjacentes (o contexto) e o cálculo final do nc-value.

O cálculo do score das palavras adjacentes começa obtendo-se uma lista de termos revisada e extraindo as palavras adjacentes dessa lista. Essa lista de termos pode ser manualmente extraída, ou para manter o método totalmente automático podemos utilizar os primeiros $n\%$ candidatos extraídos pelo c-value. Nesse trabalho escolhemos manter o método totalmente automático e assim como os autores, utilizar a informação de contexto dos 5% primeiros candidatos extraídos pelo c-value. A partir dessa lista de candidatos, obtemos as palavras de contexto deles e calculamos o score para as palavras de contexto segundo a equação 3.2.

O contexto de cada termo é extraído durante a leitura do corpus, a cada n-grama lido, o contexto dele também é salvo. O contexto de um termo consiste em um janela de n palavras para a esquerda e para a direita. Nesse trabalho utilizamos $n = 3$.

Após calculado o peso de cada palavra de contexto dos termos é calculado o nc-value segundo a equação 3.3.

5.2 Alinhamento de Termos

Nesse trabalho utilizamos 3 métodos de alinhamento diferentes: o mais simples deles, o **método composicional** (MORIN; DAILLE, 2012), o método de **alinhamento de vetores de contexto**, proposto por Fung (1998) para palavras simples e aplicado para termos multipalavra em diversos trabalhos (DÉJEAN; GAUSSIER; SADAT, 2002) (DAILLE; MORIN, 2005), e o **método composicional com projeção de contexto**, que é uma extensão do método composicional.

5.2.1 Método Composicional

O método de alinhamento consiste em traduzir cada uma das palavras do termo multipalavra da língua origem usando um dicionário bilíngue e buscar o resultado dessa tradução na lista de termos da língua alvo.

Como estamos tratando línguas com características diferentes (por exemplo, no Inglês os adjetivos geralmente vem antes do substantivo e no Português os adjetivos vem depois) ao procurar pela tradução de um termo na lista da língua alvo temos que procurar todas as permutações possíveis das palavras envolvidas. Exemplificando esta situação: o termo *annual meeting* em inglês tem traduções para ambas suas palavras no dicionário bilíngue utilizado, *annual* é traduzido para "anual" e *meeting* para "encontro". Se usarmos apenas a ordem em que as palavras aparecem na língua origem teremos "anual encontro", que não existe nos termos encontrados da língua alvo, mas se procurarmos também pelas outras ordenações possíveis das palavras, teremos "encontro anual" que é um termo válido na língua alvo. Assim temos que para um termo com n palavras (distintas) serão geradas $n!$ possíveis permutações dessas palavras.

Porém ainda temos outra variável a ser levada em conta no momento em que geramos todas as possíveis permutações de um termo: o número de traduções possíveis para uma palavra no dicionário. A palavra *free*, por exemplo, pode ser traduzida como "livre" ou "gratuito", no termo *free access* temos que gerar todas as permutações possíveis levando em conta as duas traduções de *free*. Assim para esse termo específico teremos, com n sendo o número de palavras do termo, $2 * n! = 4$ permutações possíveis, sendo elas: "acesso livre", "acesso gratuito", "livre acesso" e "gratuito acesso", onde somente as duas primeiras encontram-se na lista de termos candidatos da língua alvo. Generalizando essa abordagem podemos ver que além de fazer as permutações de cada palavra devemos multiplicar pelo número de traduções encontradas no dicionário para cada palavra. Assim o número de possibilidades é calculado por $(\prod_{i=1}^{n} t_i)n!$, onde n é o número de palavras do termo e t_i é o número de traduções dadas pelo dicionário para a palavra i que compõe o termo.

Após a geração de todos os candidatos possíveis, ordenamos todos aqueles que foram encontrados na lista de termos da língua alvo pela sua frequência.

A vantagem desse método é que ele encontra equivalências confiáveis para termos composicionais, porém não consegue encontrar nenhuma equivalência para termos não composicionais como "faculdade de direito" (*law school*) e "fuso horário" (*time zone*). Outra desvantagem é que ele é totalmente dependente da qualidade do dicionário bilíngue utilizado, caso qualquer uma das palavras que compõem o termo não estejam no dicionário não será possível gerar um candidato para ele. Dessa maneira muitos termos ficam sem nenhum candidato a equivalente.

5.2.2 Alinhamento de Vetores de Contexto

O método de Alinhamento de Vetores de Contexto (ver seção 3.2.2) utiliza as palavras que ocorrem ao redor de um termo para construir um vetor de palavras (o vetor de contexto) que caracterize-o. Os vetores de cada termo são comparados para gerar um escore de similaridade que indicará os termos equivalentes. Esse método de alinhamento é feito em três etapas: a **extração dos termos**, a **extração do contexto** e o **cálculo da similaridade** entre os termos. Por fim, na seção 5.2.2.3 detalharemos o processo de indexação do corpus, na seção 5.2.2.4 explicaremos a utilização dos dicionários bilíngues durante a tradução dos vetores e na seção 5.2.2.5 mostramos um exemplo completo da extração de um vetor de contexto e a comparação dele com outro vetor.

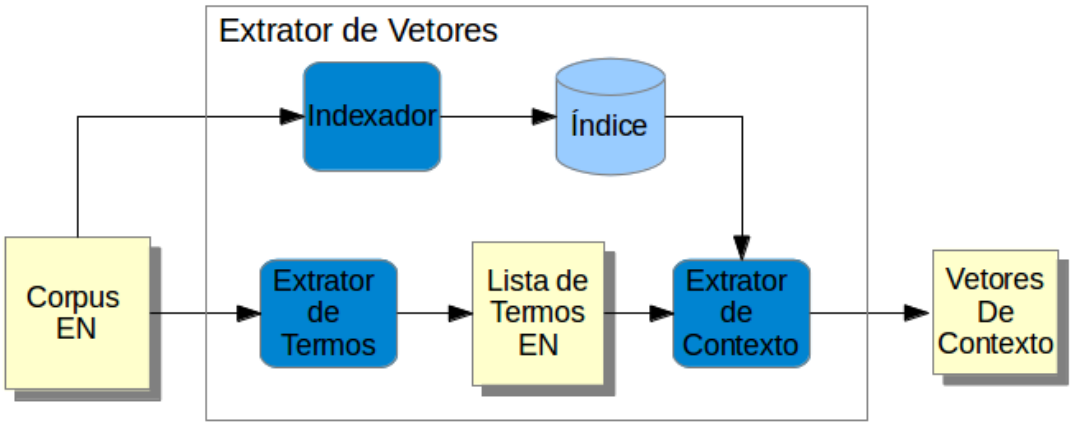

Figura 5.2 – *Pipeline* de Extração de Vetores de Contexto

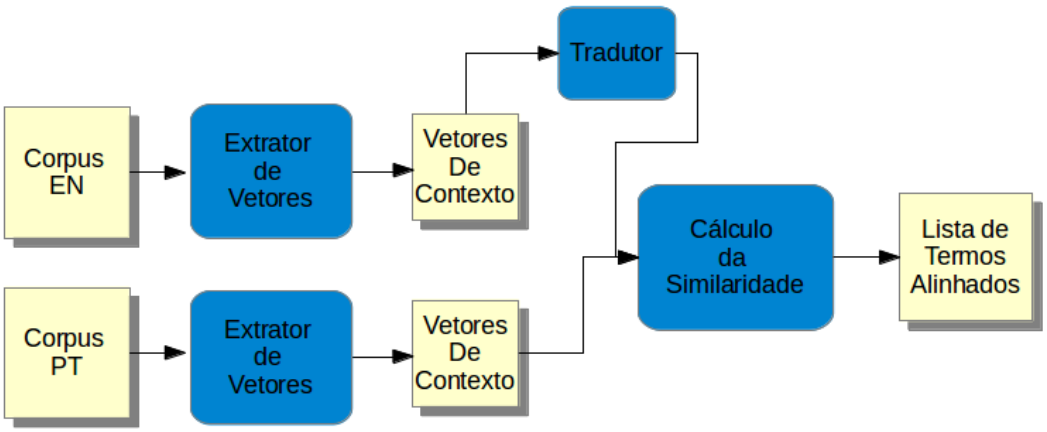

Figura 5.3 – *Pipeline* de Alinhamento de Vetores de Contexto

Na figura 5.2 temos as primeiras duas etapas, que são comuns para ambas as línguas e

podemos ver o *pipeline* completo na figura 5.3 com a última etapa, o Cálculo da Similaridade que é feito utilizando os resultados de ambos os corpora. Também podemos ver a entrada e os resultados intermediários de cada etapa do processamento.

A extração de termos funciona como explicado na seção 5.1, porém para o alinhamento utilizamos dois corpora, um em português e outro em inglês, e extraímos uma lista de termos multipalavra de ambos os corpora. Essas duas listas são a entrada para a próxima etapa, a construção dos vetores de contexto.

5.2.2.1 Extração do Contexto

A segunda etapa no Alinhamento de Vetores de Contexto, a extração do contexto começa com a obtenção dos n primeiros termos de cada uma das línguas, utilizando uma das métricas descritas na seção 5.1.6. Essa restrição de se trabalhar com apenas os n primeiros termos no ordenamento de cada língua é causada pelo custo de processamento e memória.

É preciso obter o contexto de cada termo em ambas as línguas, isto é, as palavras que ocorrem junto com eles. Para isto consultamos o índice previamente construído (ver seção 5.2.2.3) e obtemos todas as frases onde o termo ocorre. A partir das frases obtemos as palavras que co-ocorrem com o termo em uma janela que pode ser medida em número de palavras ou em número de frases ao redor do termo em questão. Também filtramos as palavras de contexto pelo seu PoS, onde esse filtro pode considerar somente substantivos, somente adjetivos, somente verbos, somente advérbios ou uma combinação dos diferentes PoS para formar um contexto. Com esse filtro queremos saber qual a classe gramatical de palavras é mais importante para a caracterização dos termos multipalavra nessa tarefa de Alinhamento multilíngue.

Durante a etapa de obtenção do contexto contamos a frequência de co-ocorrência das palavras de contexto com cada termo em questão e usamos esta informação para o cálculo da medida de associação entre o termo e a palavra de contexto. Essa medida de associação é usada para ordenar as palavras de contexto e manter somente as n com maior escore, que serão usadas para a comparação dos vetores. Foram testadas (em execuções completas do sistema) como medidas de associação o PMI (Pointwise Mutual Information) e a frequência, e os melhores resultados foram obtidos com a frequência, que são apresentados no capítulo 6. As palavras de contexto selecionadas com o PMI incluem palavras raras, pois o PMI prefere aquelas com baixa frequência. Isso impacta no cálculo de similaridade entre os vetores, porque as palavras escolhidas pelo PMI não são facilmente encontradas nos outros vetores. O aspecto positivo dos melhores resultados serem obtidos com a frequência é que ela é mais simples e fácil de calcular do que o PMI. Ao final da extração do contexto temos, para cada termo, uma lista de n palavras e sua associação.

5.2.2.2 *Cálculo da Similaridade*

A última etapa do Alinhamento de Termos é o Cálculo de Similaridade entre os termos mutipalavra extraídos em cada uma das línguas. Para cada termo multipalavra e seu contexto, calculamos a similaridade dele com todos os termos multipalavra extraídos na outra língua, gerando assim, uma lista ordenada de candidatos a tradução para cada termo da língua origem.

A similaridade é calculada utilizando similaridade de cossenos, segundo a equação 5.1, onde A e B são vetores de contexto, e cada elemento A_i ou B_i desse vetor é a medida de associação da palavra i com o termo que o vetor representa.

$$similaridade(A, B) = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{5.1}$$

A similaridade de cossenos compara a orientação dos vetores, onde caso os vetores sejam idênticos (em orientação e não necessariamente em magnitude) o valor da similaridade será 1, e caso os vetores sejam completamente opostos (formam um ângulo de 180^o), a similaridade será -1 (o que não ocorre no nosso caso, pois nenhum elemento dos vetores será negativo para que os vetores sejam opostos). Caso os vetores formem um ângulo de 90^o, a similaridade será 0 (que é o menor valor possível no nosso experimento, significando que os vetores não tem nenhuma palavra em comum).

5.2.2.3 *Indexação do Corpus*

Para um corpus de tamanho médio como o Cameleon (aproximadamente 16 milhões de palavras) temos que utilizar uma abordagem eficiente para localizar a frase em que os termos se encontram e assim extrair as palavras que ocorrem junto a cada termo.

A abordagem mais adequada encontrada foi a de indexar o corpus para que fosse possível encontrar todas as ocorrências de um termo rapidamente e assim obter as palavras que ocorrem junto com o termo para a construção dos vetores de contexto. Adicionalmente, como queremos encontrar cada frase onde ocorre o termo, precisamos indexar o corpus a nível de frase, para que o resultado de uma busca nos retorne em quais frases o termo está presente e não em qual arquivo, como é geralmente feito. Para dividir o corpus em frases utilizamos o modelo de detecção de sentenças do OpenNLP[4]. Durante a divisão do corpus em frases, criamos um arquivo diferente para cada frase tendo como nome o número do arquivo e da frase que ele contém, assim podemos acessar a frase onde ocorre o termo que buscamos e também as frases anteriores e posteriores somente modificando o nome do arquivo que queremos acessar.

Como exemplo (ver figura 5.4), se temos um corpus composto por dois textos ou arquivos, o primeiro composto três frases e o segundo por duas, eles serão divididos nos seguintes arquivos: a01f01 (arquivo 1, frase 1), a01f02 (arquivo 1, frase 2), a01f03, a02f01 e a02f02. Se buscarmos

[4]Disponível em <http://opennlp.sourceforge.net/models-1.5/>

um termo e obtivermos como resultado que ele está nos arquivos a01f02 (ocorrência 1) e a02f01 (ocorrência 2), podemos acessar diretamente os arquivos que contém as frases com o termo desejado, assim como as frases adjacentes anteriores e posteriores, no caso a01f01 (anterior ocorrência 1), a01f03 (posterior ocorrência 1) e a02f01 (anterior ocorrência 2).

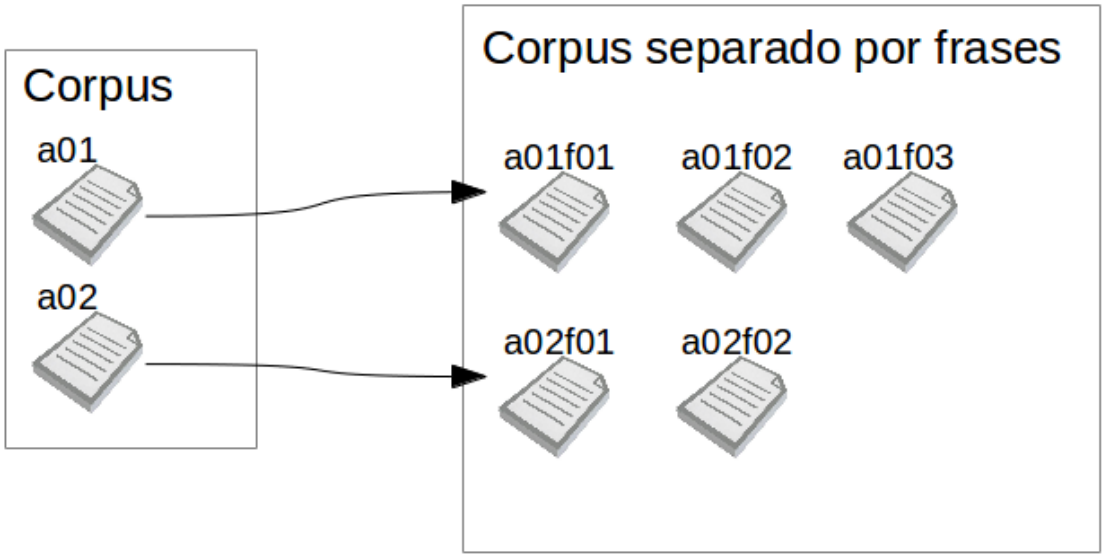

Figura 5.4 – Separação do Corpus em frases, onde a representa arquivo e f frase

Depois da separação em frases iniciamos a indexação, que foi feita utilizando o Lucene (ver seção 4.2.3), como queremos encontrar as ocorrências exatas dos termos, utilizamos o *SimpleAnalyser*, que faz somente a tokenização e passa para *lowercase* o texto.

O processo de indexação pode ser feito independentemente do processamento principal do alinhamento de termos já que diversos experimentos foram feitos com o mesmo corpus, somente com variação nos parâmetros.

5.2.2.4 Dicionários Bilíngues Utilizados

Para a tradução das palavras contidas nos vetores de contexto dos termos multipalavra e sua posterior comparação com os vetores de contexto na outra língua foram utilizados dicionários bilíngues extraídos automaticamente da web[5]. Como a cobertura desses dicionários é muito pequena, com somente cerca de 30% das palavras do corpus do Cameleon sendo traduzidas, foi necessário buscar outra fonte para complementar esse dicionário, para isso foi utilizado o tradutor da Microsoft[6] para buscar as palavras que não são encontradas nos dicionários.

Utilizando essa abordagem para traduzir os vetores de contexto podemos ter mais de uma tradução para cada palavra. Por exemplo a palavra *date* pode ser traduzida para o português como “data” ou “encontro”, ambas são traduções corretas, mas dependem da frase onde ela

[5]Disponíveis em <http://www.dicts.info/uddl.php>
[6]Disponível em <http://www.microsofttranslator.com/dev/>

está inserida. Assim, cada palavra no vetor de contexto pode ser traduzida para uma ou mais palavras e todas elas são inseridas no vetor de contexto traduzido do termo.

5.2.2.5 Exemplo

Aqui descrevemos um exemplo de como é feito o alinhamento de termos. Vamos começar a partir de um termo selecionado e verificar o conteúdo de seu vetor de contexto e a sua similaridade com o vetor de contexto do termo correspondente. Como exemplo escolhemos o termo "*law school*". Após coletar todos os termos que aparecem na janela de contexto de "*law school*", selecionamos somente as 100 primeiras palavras de contexto com o maior score de associação, representado pela frequência da palavra no corpus. Na tabela 5.3 temos as palavras com maior score

Tabela 5.3 – Vetor de Contexto de law school

Law School	
palavra	**frequência**
to	393
university	381
be	334
student	166
conference	148
college	119
program	113
professor	106
legal	105
center	98

A próxima etapa é traduzir essas palavras de contexto da língua de origem, neste caso Inglês, para a língua destino, neste caso Português usando o dicionário bilíngue. Na tabela 5.4 é possível ver o procedimento de tradução de algumas palavras do vetor de contexto.

Após a tradução do vetor, calculamos a similaridade entre todos os vetores de contexto da língua de origem com todos os vetores de contexto da língua alvo. Como exemplo mostramos o cálculo de similaridade entre o vetor traduzido de law school e o vetor de "faculdade de direito", por questões de espaço mostraremos apenas o cálculo da similaridade entre dois vetores simplificados, que podem ser vistos na tabela 5.5

Conforme visto na equação 5.1, para o cálculo da similaridade de cossenos as palavras em comum entre os vetores é que contribuem com um valor de similaridade para o numerador, pois para as palavras que não tem correspondente no outro vetor, assumimos que seu score de associação é 0, assim, no exemplo da tabela 5.5 temos 4 palavras nesse caso: universidade, conferência, professor e centro. Calculando o numerador da equação 5.1 usando a tabela 5.5 chegamos a 180114, e dividindo pelo denominador, que é o produto das magnitudes de cada

Tabela 5.4 – Tradução do Vetor de Contexto de law school

Law School	
palavra	**traduções**
to	sobre
	a
	para
university	universidade
be	existir
	viver
	ser
student	estudante
	aluno
conference	conferência
college	colégio
program	plano
	programa
legal	legal
professor	professor
center	meio
	centro
	essência

Tabela 5.5 – Vetores de contexto que serão comparados

Law School		**Faculdade de Direito**	
palavra	**frequência**	**palavra**	**frequência**
sobre	393	*universidade*	352
para	393	*professor*	238
a	393	são	200
universidade	381	paulo	151
existir	334	advogado	144
viver	334	federal	134
ser	334	curso	133
estudante	166	penal	125
aluno	166	*conferência*	96
conferência	148	mestrado	81
colégio	119	rever	78
plano	113	tema	76
programa	113	instituto	72
professor	106	lisboa	70
legal	105	internacional	69
essência	98	*centro*	67
centro	98	nacional	67

vetor, temos a similaridade entre "*law school*" e "faculdade de direito" que é igual a $0,285$. Esse valor de similaridade será usado para ordenar os candidatos a equivalentes a "*law school*".

Nesse exemplo ilustrativo calculamos o valor de similaridade usando somente as 10 primeiras palavras do vetor de contexto de contexto de "*law school*", porém utilizando o vetor de tamanho 100 definido neste trabalho o valor final de similaridade é igual a $0,358$ e "faculdade de direito" fica ranqueada em 10º lugar entre os termos candidatos a tradução, atrás de outros termos como "direito constitucional", "direito administrativo" e "bacharel em direito", todos com valores de similaridade muito próximos.

5.2.3 Método Composicional com Projeção de Contexto

O método composicional com projeção de contexto (MORIN; DAILLE, 2012) é uma extensão do método composicional para tratar os problemas que ele possui, principalmente a falta de palavras no dicionário. Para isso ele usa o contexto de uma maneira semelhante a o método de Alinhamento de Vetores. A principal diferença desse método é que ao invés de extrair o contexto para os termos completos, o contexto é extraído para cada palavra que compõe o termo individualmente. Por exemplo, ao invés de extraírmos o contexto de *"time zone"*, extraímos o contexto de *time* e o contexto de *zone*.

Este método segue as mesmas etapas básicas do método de Alinhamento de Vetores de Contexto: a extração dos termos, a extração do contexto e o cálculo de similaridade. A primeira etapa, a extração dos termos, é idêntica, já a segunda etapa, a extração do contexto é feita individualmente para cada palavra, dependendo da cobertura do dicionário bilíngue utilizado:

1. Caso a palavra esteja no dicionário o contexto da tradução é buscado no corpus da língua alvo e é criado um vetor para cada possível tradução da palavra.
2. Se a palavra não está no dicionário o contexto é buscado no corpus da língua origem e as palavras do contexto são traduzidas utilizando o dicionário. Todas as possíveis traduções são inseridas no vetor e caso alguma palavra não tenha tradução no dicionário, ela é descartada.

Ao final da extração de contexto, para um termo composto de n palavras, teremos n listas de vetores de contexto, cada lista contendo 1 ou mais vetores representando aquela palavra.

A terceira e última etapa, o cálculo de similaridade é feita entre as n listas de vetores de contexto dos termos da língua origem e os m vetores de contexto dos termos da língua destino. Ao contrário do método de Alinhamento de Vetores de Contexto em que há um vetor para cada termo e é calculado apenas um score de similaridade, nesse método são consideradas todas as possibilidades, ou seja, calculamos a similaridade entre os vetores que representam cada par de palavras possível de ser formado, com um elemento desse par pertencendo ao termo da língua origem e o outro elemento à língua alvo. Dessa forma temos $(\prod_{i=1}^{n} t_i)n!$ formas de se calcular a similaridade entre dois termos, onde $n!$ se refere a quantidade de maneiras que podemos

criar pares distintos usando todas as palavras de ambos os termos e t_i é o número de traduções que a palavra i possui no dicionário, que também é o tamanho da lista de vetores de contexto produzida para aquele termo. Para cada maneira de combinar as palavras, teremos n pares de palavras que produzirão um valor de similaridade. Calculamos a média geométrica dessas similaridades e dentre todas as médias, escolhemos a maior delas como similaridade entre os termos envolvidos.

Como exemplo concreto dessa maneira de combinar os palavras para produzir os scores de similaridade queremos calcular a similaridade entre "*date format*" e "formato da data". No dicionário utilizado encontramos duas traduções para "*date*", "encontro" e "data" e uma para "*format*", "formato" e para cada uma dessas palavras temos um vetor de contexto associado, bem como para cada palavra do termo na língua alvo, "formato da data". Nessas condições nós podemos formar $3! * 2 * 1 = 12$ listas de pares para o cálculo da similaridade, onde o $3!$ se refere ao número de listas de pares que podem ser formados, 2 é o número de traduções de "*date*" e 1 é o número de traduções de "*format*".

Para cada par teremos um valor de similaridade entre vetores, no nosso caso, similaridade de cossenos, e para cada lista de pares calculamos a média geométrica dessas similaridades (MORIN; DAILLE, 2012). De todos esses valores de similaridade calculados, escolhemos o maior entre eles para representar a similaridade entre os termos "*date format*" e "formato da data".

Esse método permite preservar as vantagens do método composicional, gerando traduções de termos composicionais com scores de similaridade iguais a 1, pois estamos comparando vetores iguais quando a palavra existe no dicionário. Além disso ele permite uma generalização muito maior pois conseguimos encontrar traduções que não estão no dicionário e também encontrar traduções que utilizam sinônimos e palavras relacionadas ao invés de somente aquelas que são traduções exatas.

5.2.4 Comparação dos métodos

Na tabela 5.6 temos um resumo dos três métodos e das etapas de cada um deles para o alinhamento de termos. Também é possível visualizar as semelhanças e diferenças entre os métodos.

Tabela 5.6 – Comparação das etapas dos métodos de alinhamento

Etapa/Método	Composicional	Alinhamento	Projeção
Extração de Termos	Etapa comum a todos métodos		
Construção de Vetores	X	Vetores extraídos da língua do termo	Vetor extraído da língua destino (se palavra no dicionário) ou da língua de origem
Tradução	Traduz cada palavra	Traduz as palavras do vetor de origem	Traduz as palavras do vetor da língua de origem
Geração das Traduções	Gera todas as possibilidades de combinações entre as palavras traduzidas	X	Gera todas as possibilidades de combinações entre os vetores de palavras
Cálculo da Similaridade	Verifica quais combinações existem na lista de termos da língua destino e escolhe a com maior frequência	Calcula a similaridade de todos os termos origem com todos termos alvo e ordena por similaridade	Calcula a similaridade de todos com todos, utilizando todas as possibilidades de combinação

6 RESULTADOS E DISCUSSÃO

Neste capítulo apresentamos os resultados dos experimentos realizados neste trabalho, divididos em duas partes: **extração monolíngue**, na qual avaliamos a qualidade dos termos extraídos de um corpus utilizando algumas métricas e **alinhamento multilíngue**, em que avaliamos a qualidade dos alinhamentos de termos em corpora comparáveis produzidos por três algoritmos de alinhamento com diferentes parâmetros.

6.1 Avaliação da Extração Monolíngue

Avaliamos o resultado do sistema de extração de termos (descrito na seção 5.1) utilizando o Corpus GENIA (seção 4.1.1) pois ele permite uma avalição completa dos resultados, já que o corpus contém todos os seus termos anotados.

6.1.1 Resultados GENIA

O sistema desenvolvido foi utilizado para extrair os termos de tamanho 2 e 3 do GENIA, e os resultados obtidos foram avaliados utilizando os termos anotados no GENIA. Ordenamos a lista de termos filtrados por PoS e para cada métrica implementada ordenamos utilizando-a e avaliamos automaticamente os primeiros n candidatos. A tabela 6.1 mostra os resultados da precisão para cada métrica ao se levar em conta o número de candidatos da coluna da esquerda.

Na figura 6.1 podemos ver o gráfico que ilustra a tabela 6.1. Nele podemos notar que as melhores medidas são o c-value (principalmente para menos de 70 termos avaliados) e o nc-value (principalmente com mais de 1000 termos avaliados). Também é possível verificar que a frequência, a medida mais básica, obtém os melhores resultados para alguns valores de termos recuperados. Os resultados das 4 medidas com melhores resultados (frequência, c-value, nc-value e poisson stirling) tem valores muito similares, e bem superiores às outras três medidas.

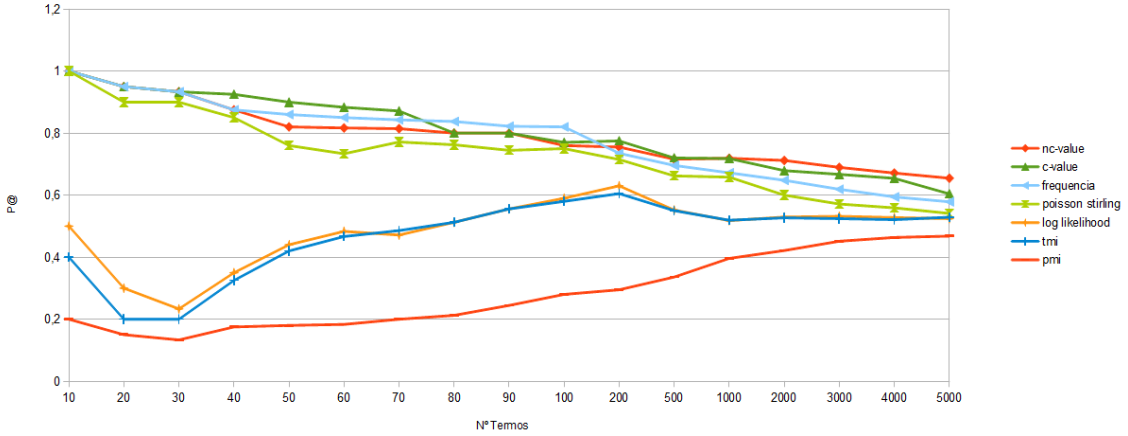

Figura 6.1 – p@10 - p@5000

Tabela 6.1 – p@n - Precisão avaliando os n primeiros resultados, onde LL = *Log Likelihood Ratio*, PMI = *Pointwise Mutual Information*, TMI = *True Mutual Information* e PS = *Poisson Stirling Measure*

n	nc-value	c-value	frequência	PS	LL	TMI	PMI
10	**1,000**	**1,000**	**1,000**	**1,000**	0,500	0,400	0,200
20	**0,950**	**0,950**	**0,950**	0,900	0,300	0,200	0,150
30	**0,933**	**0,933**	**0,933**	0,900	0,233	0,200	0,133
40	0,875	**0,925**	0,875	0,850	0,350	0,325	0,175
50	0,820	**0,900**	0,860	0,760	0,440	0,420	0,180
60	0,817	**0,883**	0,850	0,733	0,483	0,467	0,183
70	0,814	**0,871**	0,843	0,771	0,471	0,486	0,200
80	0,800	0,800	**0,838**	0,763	0,513	0,513	0,213
90	0,800	0,800	**0,822**	0,744	0,556	0,556	0,244
100	0,760	0,770	**0,820**	0,750	0,590	0,580	0,280
200	0,755	**0,775**	0,735	0,715	0,630	0,605	0,295
500	0,716	**0,720**	0,696	0,662	0,552	0,550	0,336
1000	**0,719**	0,718	0,672	0,658	0,518	0,519	0,396
2000	**0,712**	0,679	0,648	0,600	0,530	0,527	0,422
3000	**0,690**	0,667	0,619	0,571	0,532	0,524	0,451
4000	**0,671**	0,654	0,594	0,559	0,528	0,521	0,464
5000	**0,655**	0,604	0,579	0,541	0,524	0,529	0,468

6.1.2 Resultados Cameleon

A extração de Termos realizada utilizando o corpus do Cameleon resultou em 36.471 termos para o Inglês e 13.793 para o Português. Como esse corpus não possui uma versão anotada, não é possível realizar uma avaliação detalhada como no GENIA, por esse motivo utilizamos esse corpus principalmente para a avaliação do Alinhamento. Mesmo assim, uma avaliação manual (ver seção 6.2.1) dos 143 primeiros pares extraídos pelo Alinhamento resultou em 131 destes marcados como termos multipalavra por pelo menos dois de três avaliadores. Isso resulta em uma precisão de 91%.

6.1.3 Discussão

Durante a execução desse trabalho também foi realizada a extração de termos em diversos outros corpora, tanto de domínio geral (CETENFolha) (BOOS; PRESTES; VILLAVICENCIO, 2014), como de domínio específico (engenharia de software, conferências, dermatologia) e pudemos notar alguns pontos importantes de serem comentados:

Erros de atribuição de PoS: O TreeTagger foi escolhido como o PoSTagger para este trabalho por ser um dos mais utilizados na literatura, dar suporte a diversas línguas e

também ser um dos mais rápidos para processamento de grandes corpora[1]. Apesar de todas as suas qualidades, o TreeTagger possui alguns pontos fracos que observamos:

(1) No processamento do português, grande parte das palavras que iniciam frases com letra maiúscula são classificadas como substantivos (NOM).

(2) Tanto no Inglês como no Português existem diversos erros notáveis: para o inglês, por exemplo, a palavra *access* deve ser classificada como verbo, como na frase "*Touch a button to access that application*", porém *access* pode ser usado de outras maneiras como no termo *access point* em que é um substantivo, como na frase "*Edit or create a new access point*". Dessa forma o TreeTagger erra em algumas das ocorrências dessa palavra, por vezes fazendo com que resultados corretos sejam perdidos (substantivo classificado como verbo) ou que resultados incorretos sejam recuperados (verbo classificado como substantivo). Outro exemplo do mesmo tipo é a palavra *operating*, parte do termo *operating system*, que mesmo em frases como "*The amount of internal phone storage used by the operating system*" é classificada como verbo.

Para o Português podemos citar como exemplos de erros de atribuição de PoS a palavra "toque", que faz parte do termo correto "tela de toque", como um substantivo, mas também é um verbo em algumas situações e as classificações erradas fazem com que termos como "toque em contato" e "toque na tecla" sejam errôneamente extraídos.

6.2 Avaliação do Alinhamento

Para a avaliação dos resultados do alinhamento o ideal é utilizar uma lista de referência com termos de domínio e suas respectivas traduções, porém esse é um recurso extremamente difícil de se obter. Assim se optou por uma avaliação manual como alternativa e foi desenvolvido um aplicativo para que especialistas pudessem julgar os alinhamentos obtidos automaticamente com os métodos descritos neste trabalho.

6.2.1 Aplicativo de Avaliação

O aplicativo foi construído para prover suporte à tarefa de avaliação dos alinhamentos que foi realizada manualmente. Ele mostra um termo na língua origem e os n (parametrizável) primeiros candidatos a ser sua tradução na metade superior da tela (figura 6.2), na metade inferior é mostrado o contexto do termo e da tradução selecionada. O contexto são as frases onde o termo ocorre, que são buscadas no mesmo índice que foi usado durante a extração dos contextos durante o Alinhamento (seção 5.2.2.3), e são exibidas para que o especialista possa examinar a fonte de onde os termos foram extraídos.

[1] É o mais rápido segundo a avaliação disponível em <http://mattwilkens.com/2008/11/08/evaluating-pos-taggers-speed/>

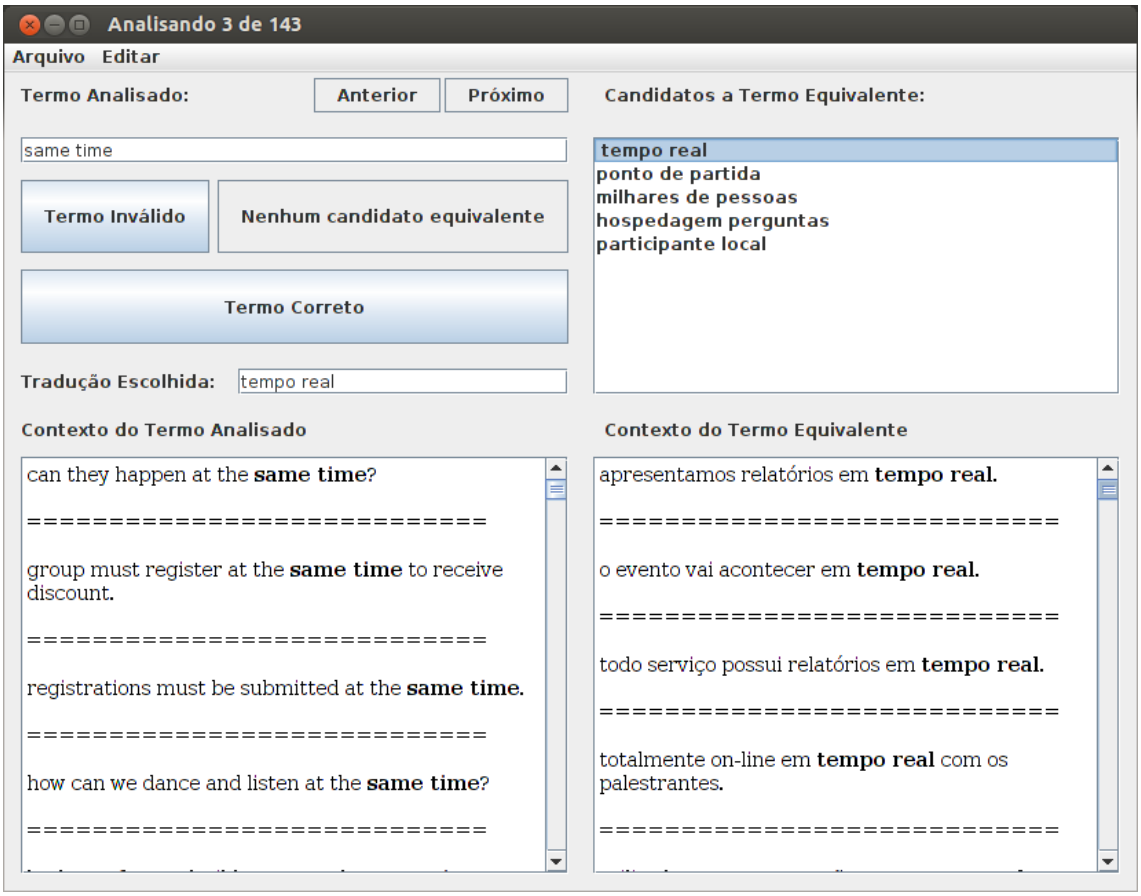

Figura 6.2 – Ferramenta para Validação dos Alinhamentos

É possível avaliar cada termo de três maneiras:

1. **Termo Inválido**: Caso o termo avaliado não seja considerado um termo multipalavra.
2. **Nenhum Candidato Equivalente**: Caso nenhuma tradução correta para o termo multipalavra esteja nos candidatos listados.
3. **Termo Correto**: O termo avaliado é um termo multipalavra válido e sua tradução foi encontrada e selecionada na lista dos candidatos.

Utilizando esta interface, foram avaliados os 143 termos com maior valor de similaridade por 3 especialistas, ao final desta avaliação, 12 desses termos foram considerados inválidos por pelo menos 2 dos avaliadores e foram descartados da avaliação do alinhamento, totalizando 131 termos válidos.

6.2.2 Experimentos e Resultados

Foram realizados diversos experimentos com os três métodos estudados: o método Composicional (5.2.1), o método de Alinhamento de Vetores de Contexto (5.2.2) e o método Composicional com projeção de Contexto (5.2.3). A etapa de Extração dos Termos é a mesma para os três métodos e os resultados foram descritos na seção 6.1.2.

6.2.2.1 Parâmetros

Com o objetivo de encontrar termos multipalavra alinhados para Inglês-Português, analisamos o impacto dos tipos de classes gramaticais (PoS tags) das palavras de contexto e do tamanho da janela de contexto na qualidade dos resultados do alinhamento.

Para o PoS queríamos analisar o quão informativo cada uma dessas classes é em caracterizar termos multipalavra e seus correspondentes na língua alvo. Para isso consideramos separadamente quatro classes e também algumas combinações delas:

- verbos (V);
- substantivos (N);
- adjetivos (J);
- advérbios (R);
- sustantivos e adjetivos (NJ);
- substantivos e verbos (NV);
- substantivos e advérbios (NR);
- substantivos, adjetivos e advérbios (NJR);
- substantivos, verbos e advérbios (NVR);
- substantivos, verbos e adjetivos (NVJ) e;
- todos os PoS - substantivos, adjetivos, advérbios e verbos.

Utilizamos 5 tamanhos de janela diferentes usando palavras ou frases ao redor do termo alvo. Esses diferentes tamanhos buscam representar uma ampla variedade de contextos desde os pequenos, com pouca informação (com somente 7 palavras), como os contextos mais intuitivos (a frase em que o termo está) até contextos grandes (7 frases ao redor do termo):

- 7 palavras (7w) onde o termo alvo é o centro da janela e utilizamos 3 palavras anteriores ao termo e 3 palavras posteriores (ver figura 6.3),

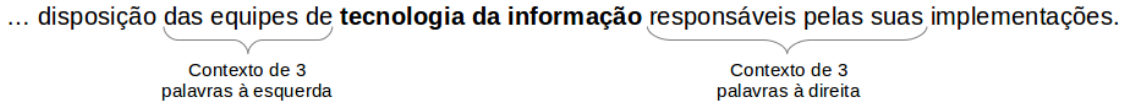

Figura 6.3 – Exemplo de janela de contexto 7w

- 21 palavras (21w) onde o termo alvo é o centro da janela e utilizamos 10 palavras anteriores ao termo e 10 palavras posteriores,
- 1 frase (1s), que é a frase que inclui o termo alvo,
- 3 frases (3s), que inclui não somente a frase com o termo alvo, mas também a anterior e a próxima frase,
- 7 frases (7s), que inclui a frase com o termo alvo e também 3 frases anteriores e 3 posteriores.

Executamos o processo de Alinhamento para cada uma das combinações de classe gramatical e tamanho da janela e avaliamos a qualidade do Alinhamento utilizando como *gold standard* os termos manualmente avaliados.

Idealmente os resultados deveriam ser analisados em termos de precisão e *recall*, mas não existe um corpus anotado com todos seus termos multipalavra para esse par de línguas. Por isso resolvemos utilizar duas medidas para a avaliação dos resultados:

1. Top$_n$, que é o percentual de termos que tiveram sua tradução encontrada entre os n primeiros candidatos.
2. MAP (*Mean Average Precision*), que leva em conta a posição em que a tradução foi encontrada. Formalmente, a MAP é calculada:

$$MAP = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{pos_i} \tag{6.1}$$

   Onde $|N|$ é o tamanho da lista de avaliação, e pos_i é a posição em que a tradução correta foi encontrada na lista ordenada de candidatos a tradução.

6.2.3 Método Composicional

Os resultados do método composicional (descrito na seção 5.2.1) foram que dentre os 131 termos analisados, encontramos os termos equivalentes de 16 deles, ou seja, um percentual de 12,21%, e um MAP de **12,21** (que é igual ao percentual pois todas as termos equivalentes encontrados estavam na 1ª posição dos termos candidatos). Esse é o método mais simples dentre os avaliados, pois somente considera as palavras encontradas no dicionário utilizado e suas possíveis permutações.

6.2.4 Alinhamento de Vetores de Contexto

Os resultados do método de Alinhamento de Vetores de Contexto (descrito na seção 5.2.2) estão nas tabelas 6.2 e 6.3 nas quais podemos ver os resultados percentuais de traduções encontradas e na tabela 6.4 onde podemos ver os resultados utilizando MAP, que possibilita uma melhor visualização da qualidade dos resultados, visto que os resultados encontrados nas primeiras posições são mais importantes e contribuem mais para o valor do MAP.

Tabela 6.2 – Resultados Percentuais com Janela de Palavras - Alinhamento de Vetores de Contexto

	7w				21w			
	Top1	Top5	Top10	Top100	Top1	Top5	Top10	Top100
V	0,00	0,00	0,00	6,87	0,00	0,00	0,00	5,34
J	3,05	6,87	9,92	33,59	3,82	6,87	9,92	32,06
R	0,00	0,76	3,05	9,16	0,00	0,76	1,53	6,87
N	9,92	22,90	28,24	51,91	6,87	18,32	25,19	**51,91**
NV	9,16	18,32	25,19	46,56	6,87	17,56	25,19	45,04
NJ	12,98	**25,19**	29,01	52,67	11,45	**24,43**	28,24	50,38
NR	8,40	22,14	25,95	51,15	6,11	19,08	24,43	50,38
NVR	9,16	19,08	26,72	47,33	6,87	17,56	25,19	45,04
NVJ	10,69	24,43	29,01	48,85	11,45	**24,43**	28,24	45,80
NJR	**13,74**	24,43	**30,53**	**52,67**	**12,21**	23,66	**30,53**	51,15
Todos	10,69	**25,19**	29,01	48,85	11,45	22,90	28,24	47,33

6.2.5 Método Composicional com Projeção de Contexto

As tabelas 6.5 e 6.6 mostram os resultados para o método Composicional com Projeção de Contexto (5.2.3). Nela podemos ver que esse método é muito mais estável, não sofrendo grandes variações de performance com a variação dos parâmetros estudados (tamanho da janela e PoS das palavras de contexto).

Tabela 6.3 – Resultados Percentuais com Janela de Frases - Alinhamento de Vetores de Contexto

	1s				3s				7s			
	Top1	Top5	Top10	Top100	Top1	Top5	Top10	Top100	Top1	Top5	Top10	Top100
V	0,00	0,00	0,00	3,82	0,00	0,00	0,76	4,58	0,00	0,00	0,00	5,34
J	6,87	12,21	17,56	30,53	7,63	14,50	18,32	29,77	6,11	12,98	17,56	32,82
R	0,00	0,76	1,53	5,34	0,00	1,53	2,29	6,87	0,76	2,29	2,29	7,63
N	16,79	29,01	36,64	52,67	15,27	25,95	34,35	51,15	16,03	25,95	32,82	52,67
NV	16,79	29,77	35,11	52,67	16,03	27,48	32,06	51,91	16,03	27,48	32,06	53,44
NJ	22,90	**38,93**	**44,27**	54,96	**22,90**	39,69	42,75	53,44	**24,43**	**39,69**	**44,27**	56,49
NR	17,56	29,01	35,11	52,67	16,03	25,95	35,11	51,15	15,27	25,95	32,82	53,44
NVR	16,03	29,77	35,11	52,67	15,27	25,95	32,06	51,91	14,50	26,72	32,82	53,44
NVJ	**25,95**	**38,93**	43,51	**55,73**	21,37	37,40	**45,04**	54,20	23,66	38,93	42,75	**58,02**
NJR	23,66	**38,93**	**44,27**	54,96	**22,90**	**40,46**	42,75	53,44	23,66	39,69	**44,27**	55,73
Todos	25,19	**38,93**	43,51	**55,73**	22,14	37,40	43,51	**54,96**	23,66	**39,69**	41,98	56,49

Tabela 6.4 – Resultados MAP em %

	7w	21w	1s	3s	7s
V	0,17	0,12	0,00	0,14	0,12
J	5,34	5,86	10,29	10,83	10,04
R	0,71	0,53	0,55	0,78	1,31
N	15,52	13,31	22,51	21,16	21,27
NV	13,81	11,79	22,67	21,87	22,16
NJ	18,47	17,82	30,64	31,00	**31,97**
NR	14,53	12,45	22,87	21,60	20,90
NVR	14,08	12,09	22,56	21,67	21,22
NVJ	17,38	17,26	**32,02**	30,04	30,85
NJR	**18,70**	**18,31**	31,06	**31,01**	31,61
Todos	17,64	17,24	31,55	30,62	30,85

Tabela 6.5 – Resultados percentuais - Método Composicional com Projeção de Contexto

	7w				1s				7s			
	Top1	Top5	Top10	Top100	Top1	Top5	Top10	Top100	Top1	Top5	Top10	Top100
V	26,72	34,35	41,22	**54,96**	**29,01**	37,40	**41,98**	51,15	**27,48**	**37,40**	39,69	50,38
J	25,19	**37,40**	38,93	46,56	27,48	33,59	35,88	43,51	26,72	35,11	36,64	44,27
R	25,95	33,59	34,35	45,80	24,43	34,35	36,64	44,27	25,19	35,11	36,64	45,04
N	**27,48**	36,64	**44,27**	54,20	28,24	**38,17**	**41,98**	**54,20**	26,72	35,88	41,98	**53,44**
Todos	26,72	34,35	41,22	**54,96**	28,24	37,40	44,27	53,44	24,43	**37,40**	**44,27**	**53,44**

Tabela 6.6 – MAP - Método Composicional com Projeção de Contexto

	7w	1s	7s
V	31,78	33,03	32,22
J	30,38	31,02	30,82
R	29,65	29,14	29,94
N	32,44	**33,30**	31,53
Todos	30,72	33,02	30,72

6.2.6 Discussão

A performance dessa tarefa pode variar bastante entre diferentes corpora utilizados, como descrito por Prochasson e Morin (2009). Por exemplo, os resultados de Rapp (1999) para o Top_1 em um corpus de Inglês com 135 milhões de palavras e um em Alemão de 163 milhões foi de 72%, enquanto, por outro lado, Chiao e Zweigenbaum (2002) usando um corpus médico Francês/Inglês com 600 mil palavras obtiveram apenas 20% de resultados corretos para o Top_1. Entretanto ambos os resultados são para alinhamento de palavras simples.

6.2.6.1 Alinhamento de Vetores de Contexto

Como podemos perceber na tabela 6.4, os melhores resultados são obtidos utilizando combinações de substantivos com outras classes gramaticais para a extração do contexto. Os resultados para verbos, advérbios e adjetivos são muito inferiores aos melhores resultados obtidos, utilizando todos os PoS, e apenas os resultados usando somente substantivos tem um resultado próximo ao do melhor. Utilizando combinações de substantivos com outras classes obtemos resultados quase sempre melhores que o obtido somente com substantivos e em alguns casos até mesmo superiores ao resultado obtido utilizando todas as classes.

Também podemos notar que mesmo aumentando o número de candidatos avaliados, a quantidade de resultados corretos não aumenta proporcionalmente, por exemplo, se observarmos a precisão com 5 candidatos, e compararmos com a de 10 candidatos, veremos um aumento de aproximadamente 6%, o que é pouco levando em conta que estamos utilizando o dobro de candidatos. Mesmo utilizando 100 candidatos, 10 vezes mais do que 10 candidatos, a precisão aumenta por volta de 16%, o que faz com que uma lista com 5 candidatos apresente um bom custo benefício.

Nas tabelas 6.4 e 6.3 podemos ver que utilizando todos os PoS, mesmo tendo mais resultados corretos no Top_{100} usando 3 frases como janela, se utilizarmos somente 1 frase o MAP é melhor, o que indica que existe um "tamanho ideal" para a janela de contexto, visto que aumentando ela a partir de certo ponto, os resultados começam a piorar, ou seja, estamos introduzindo ruído, palavras não relacionadas ao termo em questão, no vetor de contexto.

Verificando mais detalhadamente os contextos percebemos que, por exemplo, os contextos extraídos utilizando somente verbos tem resultados muito inferiores devido a muitas falhas no tagger, especialmente para o português, onde muitas palavras, especialmente substantivos são classificados como verbos. Analisando a saída do tagger é possível perceber que essas classificações erradas são em sua maioria feitas em frases que são enumerações, listas ou menus, ou seja, frases que não possuem verbos. Esse é um problema difícil de contornar em um corpus coletado automaticamente da web, pois esse tipo de texto irá ser coletado, e como é um texto bem formado e faz parte da página, não será excluído pelo processo de filtragem durante a coleta. O uso de um parser poderia minimizar este problema, mas o processo se torna computacional-

mente mais caro e o parser também irá introduzir outros tipos de erro.

Na tabela 6.7 podemos ver claramente as classificações erradas do tagger para as dez palavras de contexto com maior frequência[2] de co-ocorrência junto com os termos ***law school*** e **faculdade de direito** extraídas utilizando somente verbos. Analisando uma a uma, percebemos que para o Inglês temos 2 palavras que não são verbos (*to* e *york*) enquanto para o Português temos uma situação oposta, somente 3 palavras são realmente verbos (são, há e foi), desse modo quando calculamos a similaridade entre esses dois vetores de contexto teremos muito poucas palavras em comum, fazendo com que a similaridade entre os vetores seja baixa.

Tabela 6.7 – 10 palavras com maior frequência do vetor de contexto, usando somente verbos, de *law school* e faculdade de direito

Law School		**Faculdade de Direito**	
palavra	**frequência**	**palavra**	**frequência**
to	393	são	200
is	133	justiça	66
has	67	praça	62
are	53	pós-graduação	54
was	52	educação	49
york	42	público	41
will	42	doutorado	40
be	38	há	36
graduated	26	foi	35
received	26	graduação	34

Ao compararmos os resultados utilizando os diferentes tipos de janela, palavras ou frases, percebemos uma diferença significativa nos resultados obtidos, principalmente para os melhores resultados (todas as classes e somente substantivos). Assim podemos identificar claramente que uma janela utilizando frases, mesmo que a separação do texto em frases seja automática e sujeita a erros, produz um contexto muito mais adequado à tarefa de alinhamento de termos multipalavra.

6.2.6.2 Método Composicional com Projeção de Contexto

No Método Composicional com Projeção de Contexto podemos perceber que para todas configurações conseguimos muitos resultados corretos na primeira posição (Top_1), tendo assim resultados muito mais homogêneos para qualquer tamanho de janela e PoS usado. Comparando com o método de alinhamento de vetores de contexto em que a precisão para o Top_1 varia entre 0 e 25%, nesse método a precisão varia somente entre 24 e 29%. A grande vantagem deste método é encontrar corretamente as traduções de termos multipalavra que podem ser obtidas através do dicionário bilíngue. Como nesse método são comparados vetores de palavras individuais, e

[2]Lembrando que utilizamos 100 palavras como contexto de cada termo e aqui estamos mostrando somente 10 para exemplificar um aspecto da extração do contexto

muitas vezes essas palavras estão no dicionário, os vetores comparados serão idênticos, pois traduzimos a palavra do termo, ao contrário do método de Alinhamento de Vetores de Contexto em que traduzimos as palavras do contexto.

Nesse método temos resultados muito mais uniformes para todas as configurações testadas, todos os resultados são próximos não tendo diferenças significativas como visto no Método de Alinhamento de Vetores de Contexto. Isso se deve ao fato de que mesmo que as palavra do contexto sejam extraídas de maneira errada devido a erros no tagger, as palavras que se encontram no dicionário terão contextos idênticos, sejam eles compostos por qualquer classe gramatical.

6.2.6.3 Significância dos resultados

O teste T de Student foi utilizado para calcular se a diferença entre os diversos resultados obtidos era significativa, calculamos a diferença de todas as configurações utilizadas entre si, mas por questões de espaço apresentamos apenas algumas delas na tabela 6.8. As configurações apresentadas na tabela foram escolhidas para cobrirem diferentes valores de MAP encontrados nos resultados.

Tabela 6.8 – Significânia da diferença entre os resultados

	7w N (15,52)	7w T (17,54)	21w N (13,31)	21w V (0,12)	1s N (22,51)	1s R (0,55)	1s V (0,00)	3s J (10,83)	3s R (0,78)	7s N (21,27)	7s R (1,31)	7s T (30,85)
7w J (5,34)	**	**	**	**	**	*	**	*	**	**	**	**
7w N				**		**	**		**	*	**	**
7w T			*	**		**	**	*	**		**	**
21w N				**	**	**	**		**	**	**	**
21w V					**			**		**		**
1s N						**	**	**	**		**	**
1s R								**		**		**
1s V								**		**		**
3s J									**	**	**	**
3s R										**		**
7s N											**	**
7S R												**

Na tabela, ** indica que os resultados são diferentes com um nível de significância de 0,01 e * indica que os resultados são diferentes com um nível de relevância de 0,05. Abaixo da sigla da configuração está o valor de MAP obtido.

6.2.6.4 Comparação com outros trabalhos

Comparando nossos melhores resultados com os trabalhos relacionados, em uma avaliação dos 100 termos multipalavra em que a tradução do termo é a tradução das partes, Daille e Morin (2005) conseguiram 89 corretos, e para 100 termos em que a tradução não é a tradução das partes, 63 corretos. Usando um método similar, em uma avaliação maior, com um total de

836 termos multipalavra, Morin e Daille(2012) conseguiram 292 traduções corretas no Top_{10}, ie, 34,9%. Nossos melhores resultados, usando três frases como janela de contexto e como PoS NVJ, conseguimos 45,04% de resultados corretos para o Top_{10}. Nossos resultados são ligeiramente melhores que os de (MORIN; DAILLE, 2012) e piores que os de (DAILLE; MORIN, 2005), essa diferença é causada principalmente pelo corpus utilizado, pois enquanto os trabalhos relacionados utilizam corpora manualmente coletado, nós usamos um corpus coletado automaticamente da web, que contém muito ruído, contendo não somente documentos exclusivamente relacionados ao domínio que queremos. Essa diferença também pode ser relacionada ao número de candidatos avaliados (menos candidatos podem levar a uma melhor acurácia, já que quanto mais termos avaliarmos, mais termos pouco frequentes teremos e mais difícil que eles estejam presentes no corpus da outra língua) e à maneira que os candidatos são escolhidos (escolhendo manualmente os candidatos sabemos que eles estarão presentes no corpus da outra língua, escolhendo automaticamente não temos garantias).

No trabalho de Hazem e Morin (2014) é feita uma comparação entre a extração de contexto baseada em modelos de janela e modelos baseados em sintaxe que usam informações de dependência obtidas por um parser. Na tabela 6.9 podemos ver os melhores resultados apresentados no artigo[3] e os melhores resultados deste trabalho, nas 4 primeiras linhas encontram-se os resultados de Hazem e Morin (2014) para os diferentes corpora utilizados e na última linha os resultados obtidos utilizando o corpus do Cameleon (seção 4.1). Na primeira coluna temos os melhores resultados obtidos utilizando janela de palavras para extração do contexto e na segunda coluna os melhores resultados utilizando outras técnicas, no caso de Hazen e Morin (2014), a combinação do modelo baseado em janela e do modelo baseado em sintaxe e no caso desse trabalho o método composicional com projeção de contexto.

Tabela 6.9 – Comparação dos resultados - MAP

	melhor resultado com janela de palavras	melhor resultado
Diabetes	20,6	27,8
Volcano	49,7	53,8
Wind Energy	29,0	34,9
Breast cancer	29,3	36,1
Cameleon	32,0	33,3

Como podemos ver, os resultados obtidos neste trabalho utilizando um corpus automaticamente extraído da web, sem nenhum tipo de revisão, para o par de línguas Português/Inglês, conseguiu atingir resultados semelhantes ao estado da arte para corpora manualmente revisados para o par de línguas Inglês/Francês. Também é importante notar que no trabalho de Hazem e Morin (2014) foram avaliados os alinhamentos de palavras simples e neste trabalho avaliamos o alinhamento de termos multipalavra.

[3] Os resultados completos estão na tabela 3.1

7 CONCLUSÃO E TRABALHOS FUTUROS

A motivação para este trabalho é a melhoria da comunicação de pessoas falantes de línguas diferentes e possibilitar o acesso das pessoas a mais informações sobre diversos assuntos visto que existe muita informação restrita a algumas línguas, principalmente na internet, e com a melhoria da tradução automática essa informação seria muito mais acessível. Um dos pontos fracos da tradução automática são as expressões multipalavra, especialmente aquelas usadas em domínios e assuntos específicos (chamadas de termos multipalavra), por isso estudamos métodos para auxiliar o processo de criação de dicionários contendo esse tipo de expressão.

Neste trabalho foram analisadas diversas técnicas de extração e de alinhamento de termos multipalavra a partir de corpora comparáveis. A extração e identificação de termos multipalavra é um tópico bastante estudado, mas ainda com muitos avanços a serem feitos pois não existe uma metodologia "infalível" nessa área e depende do corpus que estamos utilizando e do objetivo que temos. Outro fator importante é a língua que estamos trabalhando, pois os recursos disponíveis para algumas poucas línguas, como o Inglês, são muito mais numerosos, variados e de maior cobertura do que os da maioria das outras, como o Português.

Tínhamos como objetivos principais neste trabalho fazer a identificação monolíngue de termos e aplicar as técnicas de alinhamento para o português, visto que não existem trabalhos que utilizem corpora comparáveis para este tipo de tarefa e avaliar os diferentes parâmetros de tamanho e tipo (PoS utilizados) de janela para a extração de contexto.

No Alinhamento de Termos, conseguimos aplicar alguns algoritmos e estudar diversos dos seus parâmetros em um corpus comparável, coletado automaticamente da web, obtendo resultados comparáveis ao estado da arte. Esses resultados mostram que um sistema de alinhamento pode ser muito valioso para o auxílio a construção de recursos léxicos como dicionários e glossários específicos de domínio, a maioria dos quais é atualmente construída totalmente a mão por especialistas, o que demanda um custo muito alto.

Mais especificamente descobrimos que os termos multipalavra são melhor caracterizados por contextos utilizando todos os PoS de palavras de conteúdo, e que um dos fatores que contribui para isso é que ainda temos muitos erros nos taggers para o Português. Se olharmos para cada PoS separadamente, os substantivos são muito mais significativos do que as outras classes. Também verificamos que os contextos extraídos com base em sentenças são melhores do que os contextos que utilizam palavras.

Como resultado deste trabalho foram gerados léxicos bílingues dos quais amostras foram validadas por especialistas e foram implementados algoritmos de extração de termos monolíngue e alinhamento multilíngue. Juntos os algoritmos implementados formam um framework para construção e comparação de contextos de palavras ou multipalavras que podem ser usados para diversas outras tarefas de processamento de linguagem como criação de thesaurus e simplificação de textos. Todos os algoritmos desenvolvidos estão disponíveis para a comunidade

acadêmica [1].

Como trabalhos futuros pretendemos utilizar informação sintática mais profunda, como relações de sujeito e objeto, obtidas através de um parser para expandir as metodologias de extração de contexto aqui apresentadas e verificar o impacto de utilizar esse tipo de relações no alinhamento de termos multilíngue.

Outro aspecto que pode ser explorado é utilizar dicionários para traduzir os vetores de contexto nos dois sentidos, ou seja, além de Inglês para Português como foi feito aqui, também traduzir os vetores de Português para Inglês. Desse modo teríamos duas listas de termos alinhados e poderíamos utilizar, por exemplo, um mecanismo de votação, para validar somente os alinhamentos que aparecem nas duas listas.

Durante este trabalho foram publicados dois trabalhos sobre a construção de corpus, o brWaC: *Identification of Multiword Expressions in the brWaC* (BOOS; PRESTES; VILLAVICENCIO, 2014) e *brWaC: A WaCky Corpus for Brazilian Portuguese* (BOOS et al., 2014). Esses trabalhos abordam a construção de um corpus muito grande coletado a partir da web contendo textos em português, a construção de corpus foi motivada pela falta de um corpus desse tamanho para nossa língua que pode ser usado para diversas tarefas dentro do processamento de linguagem natural. Também abordamos a extração de termos dentro desse corpus coletado como uma forma de verificar a qualidade e variedade dos textos coletados a partir da web comparando com os resultados de um corpus composto por artigos de jornais.

[1] Código disponível em <https://code.google.com/p/extracao-termos/>

REFERÊNCIAS

ANDERSEN, G. **Exploring Newspaper Language**: using the web to create and investigate a large corpus of modern norwegian. [S.l.] John Benjamins Pub.Co., 2012.

BANERJEE, S.; PEDERSEN, T. The design, implementation, and use of the Ngram Statistic Package. In: PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 2003...**Proceedings**, Mexico City: [s.n.], 2003. p. 370-381.

BASILI, R. et al. A contrastive approach to term extraction. In: PROCEEDINGS OF THE 4TH TERMINOLOGY AND ARTIFICIAL INTELLIGENCE CONFERENCE (TIA), 2001...**Proceedings**, France: [s.n.], 2001. p.119-128

BOOS, R.; PRESTES, K.; VILLAVICENCIO, A. Identification of Multiword Expressions in the brWaC. In: PROCEEDINGS OF THE NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 2014...**Proceedings**, Reykjavik: [s.n.], 2014

BOOS, R. et al. brWaC: A WaCky Corpus for Brazilian Portuguese. In: BAPTISTA, J. et al. Computational Processing of the Portuguese Language. [S.l.]: Springer, 2014. Lecture Notes in Computer Science, 8775.

CABRÉ, M.; SAGER, J. **Terminology**: Theory, Methods, and Applications. [S.l.] John Benjamins Pub.Co., 1999.

CALEFATO, F. et al. Assessing the impact of real-time machine translation on requirements meetings: a replicated experiment. In: PROCEEDINGS OF THE ACM-IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL DOFTWARE ENGINEERING AND MEASUREMENT, 2012...**Proceedings**, New York: ACM, 2012. p. 251-260.

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

CHIAO, Y.-C.; ZWEIGENBAUM, P. Looking for candidate translational equivalents in specialized, comparable corpora. In: PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS - VOLUME 2, 2002...**Proceedings**, Stroudsburg: ACL, 2002. p.1-5.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 16, n. 1, p. 22–29, mar. 1990. ISSN 0891-2017.

CILIBRASI, R. L.; VITANYI, P. M. B. The google similarity distance. **IEEE Trans. on Knowl. and Data Eng.**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 19, n. 3, p. 370–383, mar. 2007. ISSN 1041-4347.

CURRAN, J. R.; MOENS, M. Improvements in automatic thesaurus extraction. In: PROCEEDINGS OF THE WORKSHOP ON UNSUPERVISED LEXICAL ACQUISITION, 2002...**Proceedings**, [S.l.: s.n.], 2002. p. 59-66.

DAILLE, B.; MORIN, E. French-english terminology extraction from comparable corpora. In: PROCEEDINGS OF THE SECOND INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2005...**Proceedings**, Berlin: Springer-Verlag, 2005. p. 707-718.

DÉJEAN, H.; GAUSSIER, É.; SADAT, F. Bilingual terminology extraction: An approach based on a multilingual thesaurus applicable to comparable corpora. In: PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS COLING, 2002...**Proceedings**, [S.l.: s.n.], 2002. p. 218-224

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B**, v. 39, n. 1, p. 1–38, 1977.

DRYMONAS, E. G. **Ontology learning from text based on multi-word term concepts: The OntoGain method**. Dissertation (Master) — Technical University of Crete, Greece, 2009.

DUNNING, T. Accurate methods for the statistics of surprise and coincidence. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 19, n. 1, p. 61–74, mar. 1993. ISSN 0891-2017.

FRANTZI, K. T.; ANANIADOU, S.; TSUJII, J. ichi. The c-value/nc-value method of automatic recognition for multi-word terms. In: PROCEEDINGS OF THE SECOND EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 1998...**Proceedings**, London: Springer-Verlag, 1998. p. 585-604.

FUNG, P. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: PROCEEDINGS OF THE THIRD CONFERENCE OF THE ASSOCIATION FOR MACHINE TRANSLATION IN THE AMERICAS ON MACHINE TRANSLATION AND THE INFORMATION SOUP, 1998...**Proceedings**, London: Springer-Verlag, 1998. p. 1-17.

GRANADA, R. et al. A comparable corpus based on aligned multilingual ontologies. In: PROCEEDINGS OF THE FIRST WORKSHOP ON MULTILINGUAL MODELING, 2012...**Proceedings**, Stroudsburg: ACL, 2012. p. 25-31.

HAZEM, A.; MORIN, E. Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models. In: **Computational Linguistics and Intelligent Text Processing**. [S.l.]: Springer, 2014. p. 310–323.

HEYLEN, K. et al. Modelling word similarity: an evaluation of automatic synonymy extraction algorithms. In: **LREC**. European Language Resources Association, 2008.

HIPPISLEY, A.; CHENG, D.; AHMAD, K. The head-modifier principle and multilingual term extraction. **Natural Language Engineering**, New York: Cambridge University Press, v. 11, n. 2, p. 129–157, jun. 2005.

HJELM, H. **Cross-language Ontology Learning: Incorporating and Exploiting Cross-language Data in the Ontology Learning Process**. Thesis (PhD) — Stockholm University, 2009.

JACKENDOFF, R. Twistin' the night away. **Language**, Linguistic Society of America, v. 73, n. 3, p. pp. 534–559, 1997. ISSN 00978507.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 2.ed. [S.l.]: Prentice Hall, 2008.

JUSTESON, J.; KATZ, S. Technical terminology: some linguistic properties and an algorithm for identification in text. **Natural Language Engineering**, p. 9–27, 1995.

KAGEURA, K.; UMINO, B. Methods of Automatic Term Recognition - A Review. **Terminology**, [S.l.]: John Benjamins Pub.Co., v. 3, n. 2, p. 259-289, 1996.

KEENAN, E.; FALTZ, L. **Boolean Semantics for Natural Language**. D. Reidel Publishing Company, 1985. (Culture, Illness, and Healing).

KOEHN, P. et al. Moses: Open source toolkit for statistical machine translation. In: PROCEEDINGS OF THE 45TH ANNUAL MEETING OF THE ACL ON THE INTERACTIVE POSTER AND DEMONSTRATION SESSIONS, 2007...**Proceedings**, Stroudsburg: ACL, 2007. p. 177-180.

KRIEGER, M. d. G.; FINATTO, M. **Introdução à terminologia: teoria e prática**. Contexto, 2004.

LIN, D. Automatic retrieval and clustering of similar words. In: PROCEEDINGS OF THE 17TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS - VOLUME 2, 1998...**Proceedings**, Stroudsburg: ACL, 1998. p. 768-774.

MAYNARD, D.; ANANIADOU, S. Term extraction using a similarity-based approach. In: **In Recent Advances in Computational Terminology. John Benjamins**. [S.l.: s.n.], 1999. p. 261–278.

MORIN, E.; DAILLE, B. Revising the compositional method for terminology acquisition from comparable corpora. In: KAY, M.; BOITET, C. (Ed.). **COLING**. [S.l.]: Indian Institute of Technology Bombay, 2012. p. 1797–1810.

MUNIZ, M. C. M. **Léxicos Computacionais: Desafios na Construção de um Léxico de Português Brasileiro.** Dissertation (Master) — Instituto de Ciências Matemáticas de São Carlos, USP, São Carlos, 2003.

MUNIZ, M. C. M. **A construção de recursos lingüístico-computacionais para o português do Brasil: o projeto de Unitex-PB.** Dissertation (Master) — Instituto de Ciências Matemáticas de São Carlos, USP, São Carlos, 2004.

NAKAGAWA, H.; MORI, T. Automatic term recognition based on statistics of compound nouns and their components. **Terminology**, v. 9, n. 2, p. 201–219, 2003.

NOLAN, C. **Batman Begins**. 2005.

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

OHTA, T.; TATEISI, Y.; KIM, J.-D. The genia corpus: An annotated research abstract corpus in molecular biology domain. In: PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON HUMAN LANGUAGE TECHNOLOGY RESEARCH, 2002...**Proceedings**, San Francisco: Morgan Kaufmann Pub.Inc., 2002. p. 82-86.

ORENGO, V.; HUYCK, C. A stemming algorithm for the portuguese language. In: STRING PROCESSING AND INFORMATION RETRIEVAL, 2001...**Proceedings**, [S.l.: s.n.], 2001. p. 186-193.

OTERO, P. Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. PROCEEDINGS OF LREC 2008 WORKSHOP ON COMPARABLE CORPORA, 2008...**Proceedings**, [S.l.: s.n.], 2008. p. 19–26.

PAPINENI, K. et al. Bleu: A method for automatic evaluation of machine translation. In: PROCEEDINGS OF THE 40TH ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2002...**Proceedings**, Stroudsburg: ACL, 2002. p. 311-318.

PEARSON, J. **Terms in context**. [S.l.]: John Benjamins Publishing, 1998.

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

PROCHASSON, E.; MORIN, E. Anchor points for bilingual extraction from small specialized comparable corpora. **TAL**, v. 50, n. 1, p. 283–304, 2009.

QUASTHOFF, U.; WOLFF, C. The poisson collocation measure and its applications. In: SECOND INTERNATIONAL WORKSHOP ON COMPUTATIONAL APPROACHES TO COLLOCATIONS, 2002...**Proceedings**, [S.l.]: IEEE, 2002

RAMISCH, C. **Multi-word terminology extraction for domain-specific documents**. Dissertation (Master) — École Nationale Supérieure d'Informatique et de Mathématiques Appliquées, Grenoble, France, jun. 2009.

RAMISCH, C. **A generic and open framework for multiword expressions treatment: from acquisition to applications**. 246 p. p. Thesis (PhD) — University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France, sep. 2012.

RAMISCH, C. et al. Picking them up and figuring them out: verb-particle constructions, noise and idiomaticity. In: PROCEEDINGS OF THE TWELFTH CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 2008...**Proceedings**, Stroudsburg: ACL, 2008. p. 49-56.

RAPP, R. Automatic identification of word translations from unrelated english and german corpora. In: PROCEEDINGS OF THE 37TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON COMPUTATIONAL LINGUISTICS, 1999...**Proceedings**, Stroudsburg: ACL, 1999. p. 519-526.

ROCHETEAU, J.; DAILLE, B. Ttc termsuite: A uima application for multilingual terminology extraction from comparable corpora. In: PROCEEDINGS OF THE 5TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2011...**Proceedings**, Chiang Mai: [s.n.], 2011. p. 9-12.

SANJUAN, E. et al. A symbolic approach to automatic multiword term structuring. **Comput. Speech Lang.**, Academic Press Ltd., London, UK, UK, v. 19, n. 4, p. 524–542, oct. 2005. ISSN 0885-2308.

SAVARY, A.; JACQUEMIN, C. Reducing information variation in text. In: RENALS, S.; GREFENSTETTE, G. (Ed.). **Text- and Speech-Triggered Information Access**. Springer Berlin Heidelberg, 2003, (Lecture Notes in Computer Science, v. 2705). p. 145–181.

SCHMID, H. **Probabilistic Part-of-Speech Tagging Using Decision Trees**. 1994.

SMADJA, F. Retrieving collocations from text: Xtract. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 19, n. 1, p. 143–177, mar. 1993. ISSN 0891-2017.

STALLONE, S. **Rocky Balboa**. 2006.

TSVETKOV, Y.; WINTNER, S. Extraction of multi-word expressions from small parallel corpora. **Natural Language Engineering**, v. 18, p. 549–573, 2012. ISSN 1469-8110.

WILKS, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. **Ann. Math. Statist.**, The Institute of Mathematical Statistics, v. 9, n. 1, p. 60–62, 03 1938.

WONG, W.; LIU, W.; BENNAMOUN, M. Determining termhood for learning domain ontologies in a probabilistic framework. In: PROCEEDINGS OF THE SIXTH AUSTRALIAN CONFERENCE ON DATA MINING AND ANALYTICS - VOLUME 70, 2007...**Proceedings**, Darlinghurst: Australian Computer Society, Inc., 2007. p. 55-63.

ZHANG, Z.; BREWSTER, C.; CIRAVEGNA, F. Ciravegna f: A comparative evaluation of term recognition algorithms. In: PROCEEDINGS OF THE SIXTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 2008...**Proceedings**, [S.l.: s.n.], 2008.