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

OTAVIO COSTA ACOSTA

Identificação e Tratamento de Expressões Multipalavras aplicado à Recuperação de Informação

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

Orientador: Profa. Dra. Aline Villavicencio
Coorientador: Profa. Dra. Viviane Moreira

Porto Alegre
2011

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Acosta, Otavio Costa
Identificação e Tratamento de Expressões Multipalavras aplicado à Recuperação de Informação / Otavio Costa Acosta. – 2011.
64 f.:il.
Orientador: Aline Villavicencio; Coorientador: Viviane Moreira.
Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2014.
1.Processamento de Linguagem Natural. 2.Expressão Multipalavra 3.Recuperação de Informação. I. Villavicencio, Aline. II. Moreira, Viviane. III. Título.

"*Keep smiling and never give up even when things get you down*."

— Gary Gordon

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todos contribuíram de alguma forma para o desenvolvimento deste trabalho:

Em especial à minha orientadora, Profa. Dra. Aline Villavicencio, por toda confiança, paciência e oportunidades dadas, sempre buscando transmitir seus conhecimentos e experiências da melhor maneira possível. Agradeço por toda orientação e atenção dispensada durante o mestrado, as quais foram de suma importância para a realização deste trabalho.

À minha coorientadora, Profa. Dra. Viviane Moreira, por sempre estar disposta a ajudar, colaborar com sugestões e comentários significativos, além da disponibilização de material utilizado no desenvolvimento deste trabalho. Obrigado por toda a atenção e paciência dispensada a mim ao longo deste período.

Ao meu pai, por todo apoio carinho, amor e compreensão, sendo a base de minha estrutura mesmo à distância. À minha mãe, mesmo não estando mais presente, com certeza sempre esteve na torcida, enviando energias positivas. Às minhas irmãs, sobrinhas e cunhado, por todo carinho e ajuda, mesmo à distância.

À minha namorada, Jessica Marques da Silva, por todo seu amor, sempre me incentivando e apoiando. Agradeço por toda confiança e incentivo dado, e também ser meu conforto nos momentos mais adversos.

Aos meus colegas de jornada: Eduardo Borges, Euler de Oliveira, Giseli Lopes, Gustavo Piltcher, Marcos Nunes, Sérgio Fujii que também são grandes amigos e compartilharam momentos de integração e descontração, assim como sempre estiveram dispostos a ajudar de diversas maneiras, em especial à Giseli e ao Marcos por todo apoio. Agradeço também a todos meus amigos que me acompanharam ao longo deste período.

Ao grupo de PLN da UFRGS, em especial ao Mário Machado, Paulo Schreiner, Rodrigo Wilkens e Carlos Ramish, por sempre estarem dispostos a discussões, sugestões e trocas de experiências que foram fundamentais para o desenvolvimento deste trabalho. Ao Leonardo Zilio, por ter colaborado com seu conhecimento de anotação para a realização de um experimento.

Ao Instituto de Informática da UFRGS, por toda infraestrutura disponibilizada e a seus profissionais, aos professores do PPGC por todo ensinamento e aos funcionários sempre atenciosos e prestativos. À CAPES, pelo apoio financeiro, permitindo minha dedicação exclusiva durante 11 meses deste trabalho.

Este trabalho foi realizado com o apoio dos projetos CAPES - COFECUB (707/11), CNPq (479824/2009-6, 202007/2010-3 e 309569/2009-5).

RESUMO

A vasta utilização de Expressões Multipalavras em textos de linguagem natural requer atenção para um estudo aprofundado neste assunto, para que posteriormente seja possível a manipulação e o tratamento, de forma robusta, deste tipo de expressão. Uma Expressão Multipalavra costuma transmitir precisamente conceitos e ideias que geralmente não podem ser expressos por apenas uma palavra e estima-se que sua frequência, em um léxico de um falante nativo, seja semelhante à quantidade de palavras simples. A maioria das aplicações reais simplesmente ignora ou lista possíveis termos compostos, porém os identifica e trata seus itens lexicais individualmente e não como uma unidade de conceito. Para o sucesso de uma aplicação de Processamento de Linguagem Natural, que envolva processamento semântico, é necessário um tratamento diferenciado para essas expressões. Com o devido tratamento, é investigada a hipótese das Expressões Multipalavras possibilitarem uma melhora nos resultados de uma aplicação, tal como os sistemas de Recuperação de Informação. Os objetivos desse trabalho estão voltados ao estudo de técnicas de descoberta automática de Expressões Multipalavras, permitindo a criação de dicionários, para fins de indexação, em um mecanismo de Recuperação de Informação. Resultados experimentais apontaram melhorias na recuperação de documentos relevantes, ao identificar Expressões Multipalavras e tratá-las como uma unidade de indexação única.

Palavras-chave: Processamento de Linguagem Natural. Expressão Multipalavra. Recuperação de Informação.

Identification and Treatment of Multiword Expressions applied to Information Retrieval

ABSTRACT

The use of Multiword Expressions (MWE) in natural language texts requires a detailed study, to further support in manipulating and processing, robustly, these kinds of expression. A MWE typically gives concepts and ideas that usually cannot be expressed by a single word and it is estimated that the number of MWEs in the lexicon of a native speaker is similar to the number of single words. Most real applications simply ignore them or create a list of compounds, treating and identifying them as isolated lexical items and not as an individual unit. For the success of a Natural Language Processing (NLP) application, involving semantic processing, adequate treatment for these expressions is required. In this work we investigate the hypothesis that an appropriate identification of Multiword Expressions provide better results in an application, such as Information Retrieval (IR). The objectives of this work are to compare techniques of MWE extraction for creating MWE dictionaries, to be used for indexing purposes in IR. Experimental results show qualitative improvements on the retrieval of relevant documents when identifying MWEs and treating them as a single indexing unit.

Keywords: Natural Language Processing. Multiword Expression. Information Retrieval.

LISTA DE FIGURAS

Figura 3.1: Modelo Básico - SRI 24
Figura 3.2: Exemplo – Modelo Booleano 27
Figura 3.3: Exemplo – Modelo Vetorial 28
Figura 3.4: Resultado de uma Consulta em um SRI 31
Figura 4.1: Estrutura dos Documentos Originais 37
Figura 4.2: Documentos com Inserção de EMs 44
Figura 4.3: Arquitetura 45
Figura 5.1: Tópico de Consulta 141 51
Figura 5.2: Tópico de Consulta 184 53

LISTA DE TABELAS

Tabela 4.1: Total de Documentos 36
Tabela 4.2: Índices 43
Tabela 4.3: Total EMs por Dicionário 43
Tabela 5.1: Resultados – Avaliação A 46
Tabela 5.2: *Baseline* x MCN 47
Tabela 5.3: *Baseline* x PCN 47
Tabela 5.4: Tópicos com Expressões Multipalavras 49
Tabela 5.5: Resultados – Avaliação B 49
Tabela 5.6: *Baseline* x CN 50
Tabela 5.7: *Baseline* x MCN 50
Tabela 5.8: *Baseline* x PCN 50
Tabela 5.9: *Baseline* x GS 50
Tabela 5.10: *Baseline* x AD 50
Tabela 5.11: *Baseline* x Manual 1 50
Tabela 5.12: *Baseline* x Manual 2 50
Tabela 5.13: Ranking Tópico 141 - *Baseline* 52
Tabela 5.14: Ranking Tópico 141 - CN 52
Tabela 5.15: Os Cinco Melhores Resultados 52
Tabela 5.16: Piores Resultados 53
Tabela 5.17: Ranking Tópico 184 - *Baseline* 54
Tabela 5.18: Ranking Tópico 184 - CN 54
Tabela 5.19: CN x MCN 55
Tabela 5.20: CN x PCN 55
Tabela 5.21: MCN x PCN 55
Tabela 5.22: MCN x AD 56
Tabela 5.23: MCN x M1 56
Tabela 5.24: MCN x M2 56
Tabela 5.25: GS x MCN 57

LISTA DE ABREVIATURAS E SIGLAS

EM	Expressão Multipalavra
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informação
SRI	Sistemas de Recuperação de Informação
IM	Informação M
χ^2	Chi-Quadrado
EP	Entropia de Permutação
BL	Baseline
CN	Composto Nominal
MCN	Melhores Compostos Nominais
PCN	Piores Compostos Nominais
GS	Gold Standard
AD	Árvore de Decisão
M1	Manual 1
M2	Manual 2

SUMÁRIO

1 INTRODUÇÃO 12

2 EXPRESSÕES MULTIPALAVRAS 16

2.1 Classificação de Expressões Multipalavras 17
2.1.1 Expressões Institucionalizadas 17
2.1.2 Expressões Lexicalizadas 17

2.2 Técnicas para Extração de Multipalavras 18
2.2.1 Informação Mútua 20
2.2.2 Chi-Quadrado 20
2.2.3 Entropia de Permutação 21
2.2.4 Dice 21
2.2.5 Teste-T 22
2.2.6 Outros métodos 22

3 RECUPERAÇÃO DE INFORMAÇÃO 23

3.1 Modelos Clássicos de Recuperação de Informação 26
3.1.1 Modelo Booleano 26
3.1.2 Modelo Vetorial 28
3.1.3 Modelo Probabilístico 29

3.2 Métricas de Avaliação em Recuperação de Informação 31

3.3 Expressões Multipalavras e a Recuperação de Informação 33

4 MATERIAIS E MÉTODOS 36

4.1 Recursos e Ferramentas 36

4.2 Inserção de Expressões Multipalavras como Termo Único 38

4.3 Dicionários de Expressões Multipalavras 39
4.3.1 Métodos Estatísticos 39
4.3.2 Gold Standard 41
4.3.3 Árvore de Decisão 41
4.3.4 Manual 42

4.4 Criação de Índices 43

4.5 Arquitetura 44

5 EXPERIMENTOS 46

5.1 Avaliação A – Conjunto Total de Tópicos 46

5.2 Avaliação B – Tópicos Modificados por Expressões Multipalavra 48

5.3 Avaliação C – Comparativo entre os índices 54

6 CONCLUSÕES E TRABALHOS FUTUROS 58

REFERÊNCIAS 60

APÊNDICE <ÍNDICES X BASELINE SEM 5%> 64

1 INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) é uma área da computação que abrange conceitos de Inteligência Artificial e de Linguística, estudando problemas de geração e compreensão automática de línguas humanas naturais. O PLN é considerado um dos grandes desafios da computação, pois envolve a união de conhecimentos linguísticos e computacionais, no intuito de criar mecanismos considerados eficientes e inteligentes na interação homem-máquina. Este tipo de solução tem sido amplamente utilizado para auxiliar na realização de tarefas de fins corporativos, governamentais e pessoais, uma vez que a comunicação eletrônica para negócios, educação, segurança e comércio tem aumentado constantemente, e para isso é necessário um meio que seja capaz de entender melhor a linguagem natural (MANNING; SHÜTZE, 1999).

A partir deste contexto, umas das motivações deste trabalho é a identificação e o tratamento apropriado de Expressões Multipalavras (EMs), em inglês *Multiword Expressions (MWEs)*, pois são formas de aumentar ainda mais a qualidade do entendimento e geração de linguagem, e de aplicações para a resolução deste tipo problema.

O termo "Expressão Multipalavra"" vem sendo utilizado para descrever uma grande quantidade de construções distintas, porém é usualmente relacionado com verbos de suporte, compostos nominais, frases institucionalizadas, entre outros. Calzolari et al. (2002) define EMs como uma sequência de palavras que atuam como uma única unidade, ou seja, um conjunto de palavras que representam um significado diferente do que o obtido caso as palavras desse conjunto sejam analisadas individualmente. Como exemplos de EMs, é possível citar:

- Verbos Frasais como "*carry up"* ou "*run away"*
- Compostos Nominais como *"bode expiatório"* ou *"aquecimento global"*
- Verbos de Suporte como *"tomar um banho"* ou *"fazer uma comida"*
- Expressões Idiomáticas como *"engolir sapo"* ou *"bater as botas"*

A natureza das Expressões Multipalavras pode ser bem variada e cada uma das diferentes classes de EMs têm características bem específicas, dificultando a implementação de um mecanismo que proporcione o tratamento unificado para sua identificação. Um simples sistema capaz de identificar palavras a partir de espaços encontrados em um texto, também

conhecido como processo de *tokenização*, é incapaz de reconhecer Expressões Multipalavras e tratá-las como uma unidade. Para que uma aplicação de PLN seja eficiente, são necessários mecanismos que sejam capazes de identificar EMs, mesmo com as diversas variações existentes, e que também possam tratar e fazer o uso delas de uma forma significativa (SAG et al., 2002) (BALDWIN et al., 2003).

De acordo com Jackendoff (1997) é estimado que o número de EMs no léxico de um falante nativo de uma língua seja da mesma ordem de magnitude do número de palavras simples. No entanto, essas proporções são provavelmente subestimadas se considerarmos a linguagem de um domínio específico no qual o vocabulário e a terminologia especializada vão ser compostos em sua maior parte por Sem, como por exemplo, "*sequenciamento de proteínas*", e que novas EMs estão constantemente sendo introduzidas na linguagem, como por exemplo "*gripe suína*" (VILLAVICENCIO et al., 2010).

As EMs são encontradas em todos os gêneros de textos e seu uso adequado está sendo alvo de estudo, tanto no meio linguístico quanto no meio computacional, devido às diversas variações de características desse tipo de expressão, as quais acabam gerando problemas para o sucesso de métodos computacionais que visam o seu tratamento. Para a linguística, o uso de Multipalavras poderia auxiliar na validação das propriedades de teorias gramaticais. Já para aplicações de PLN, como a tradução automática, o reconhecimento de EMs é necessário para preservar o significado e gerar traduções apropriadas, evitando a geração de traduções que não tenham sentido para o idioma correspondente (VILLAVICENCIO et al., 2005). Para exemplificar, a tradução da expressão *"kick the bucket"*, que na língua inglesa tem o significado de "*morrer*", em uma tradução literal para a língua portuguesa seria *"chutar o balde"*, o que alteraria totalmente o real sentido da expressão.

Outra área da Computação também ligada a tecnologias da linguagem e que pode se beneficiar dos avanços em PLN é a Recuperação de Informação (RI). A RI é a ciência que estuda a forma de representação, armazenamento, organização e acesso à informação, objetivando um fácil acesso aos dados que sejam de interesse do usuário (BAEZA-YATES; RIBEIRO-NETO, 1999). Segundo Sparck Jones (1997), o Processamento de Linguagem não é vital a um sistema de Recuperação de Informação moderno, mas sim conveniente. Porém, a utilização do Processamento de Linguagem Natural para a Recuperação de Informação é uma área onde ainda existe muito a ser investigado e é nisto que se baseia este trabalho.

O PLN poderia contribuir, por exemplo, na seleção de termos compostos para indexação. A seleção de termos de indexação adequados é crucial para a melhoria de qualidade dos sistemas de RI. Em um sistema ideal, os termos de indexação deveriam

corresponder aos conceitos presentes nos documentos. Se a indexação for feita somente com os termos atômicos pode haver uma perda semântica do conteúdo dos documentos. Por exemplo, ao separarmos o termo composto *"educação a distância"* em seus termos individuais estaremos perdendo seu real significado, sendo assim, torna-se necessário identificar tais termos a fim de que o sistema de RI os trate de maneira correta. Estudos comprovam que o tratamento adequado de termos compostos gera uma melhoria significativa na precisão destes sistemas. Evans e Zhai (1996) reportam melhorias de 13% para os cinco primeiros documentos recuperados. Portanto, as EMs devem ser identificadas e tratadas adequadamente, para um processamento mais preciso da linguagem e consequente melhoria da qualidade dos sistemas, especialmente para tarefas de PLN que envolvam algum tipo de processamento semântico.

Muitos sistemas de RI utilizam métodos estatísticos tradicionais para esta identificação que não fazem o uso de PLN. Nestes sistemas, as palavras que compõem um termo composto devem ocorrer com certa frequência em um contexto comum, sendo calculada a co-ocorrência entre esses pares de palavras e caso ela seja menor que um limiar determinado, o par é descartado (SALTON; MCGILL, 1983). A identificação de termos compostos representa uma oportunidade de se aplicar técnicas do Processamento de Linguagem Natural em RI. Com a devida identificação e tratamento das Multipalavras estima-se que seja possível gerar a indexação de melhores termos para um sistema que gerencie a recuperação de informação, com uma consequente melhora nos resultados obtidos.

O foco deste trabalho é voltado para a utilização de métodos de Processamento de Linguagem Natural na identificação de Expressões Multipalavras em coleções de textos (corpora), para que posteriormente, essas EMs possam ser utilizadas como unidades únicas de indexação, em um sistema de Recuperação de Informação. Assim, os resultados de consultas obtidos são comparados com base em valores dados por métricas de avaliação de RI, com o objetivo de investigar os efeitos, melhoras ou pioras que a utilização de Expressões Multipalavras ocasiona em sistemas de Recuperação de Informação.

O conteúdo deste trabalho está estruturado da seguinte forma:

- O Capítulo 2 apresenta uma visão geral sobre as Expressões Multipalavras, bem como seus tipos e alguns métodos para a extração automática em textos.
- O Capítulo 3 é voltado à Recuperação de Informação e seu uso com o Processamento de Linguagem Natural.
- No Capítulo 4 são apresentadas quais ferramentas e técnicas foram utilizadas para o desenvolvimento deste trabalho.

- No Capítulo 5 é apresentado o experimento proposto com as devidas avaliações de seus resultados.
- Por fim, no Capítulo 6 são descritas as considerações finais deste trabalho com as conclusões obtidas e com algumas perspectivas de trabalhos futuros.

2 EXPRESSÕES MULTIPALAVRAS

O termo "Expressão Multipalavra", também utilizado em inglês como *Multiword Expression*, tem sido definido de maneira distinta por pesquisadores de diferentes áreas. Abaixo, são apresentadas duas definições dadas por diferentes autores. Eles definem uma Expressão Multipalavra por:

> Uma sequência de palavras que atuam como uma simples unidade em algum nível da análise linguística. (CALZOLARI et al., 2002)

> Uma interpretação idiossincrática que ultrapassa os limites de uma palavra. (SAG et al., 2002)

Em ambas as definições, o foco é o desencontro entre a interpretação de uma EM como um todo e o significado isolado de cada palavra que compõe a expressão. Além disso, as EMs são classificadas de várias formas devido à sua grande heterogeneidade. Por exemplo, Calzolari et al. (2002) classificam as Expressões Multipalavras como expressões fixas ou semi-fixas. Entretanto, as EMs ainda podem ser separadas por tipos, como expressões idiomáticas, compostos nominais, nomes próprios, construções verbo partícula, verbos de suporte, verbos frasais, colocações, entre outras.

Um dos grandes desafios da área de PLN é a identificação dessas expressões, "escondidas" em textos de diversos gêneros. As dificuldades encontradas na identificação de Expressões Multipalavras são dadas pelos seguintes motivos:

- A dificuldade em encontrar os limites de uma Multipalavra, por ela variar na quantidade de palavras ou até mesmo por não obedecer sempre a mesma sequência. Por exemplo: *Banco de Dados Relacional* x *Banco de Dados* x *Banco* x *Dados Relacional* x *Dados* x etc.

- Em uma perspectiva multilíngue, não necessariamente as EMs de um idioma de origem possuem equivalentes no idioma de destino, ou mesmo uma palavra pode ser traduzida para uma EM. Por exemplo: a palavra "sinaleira" em português tem como equivalente *traffic lights*, em inglês.

2.1 Classificação de Expressões Multipalavras

As Expressões Multipalavras consistem em uma classe heterogênea de fenômenos, fazendo com que seja necessário classificá-las em subconjuntos menores. Isso se deve ao fato da dificuldade de se atribuir um conjunto de características que dê cobertura a toda classe das Expressões Multipalavras. É descrita a seguir, uma visão geral dos diferentes tipos de EMs, sendo adotada a classificação e terminologia utilizada por Sag et al. (2002). O primeiro tipo de divisão consiste em duas classes: Expressões Institucionalizadas e Expressões Lexicalizadas.

2.1.1 Expressões Institucionalizadas

As Expressões Institucionalizadas *(Institutionalised Phrases)* referem-se à EMs que são sintática e semanticamente composicionais porém, sua co-ocorrência é convencionalizada. Por exemplo, "*strong tea*" é formado por duas palavras combinadas que ocorrem em uma frequência grande comparado a possíveis "*sinônimos*" do tipo "*powerful tea*" ou "*potent tea*", que são encontradas com frequências nulas ou baixas em relação à EM citada anteriormente.

2.1.2 Expressões Lexicalizadas

As Expressões Lexicalizadas (*Lexicalised Phrases*) correspondem a expressões que são sintática ou semanticamente idiossincráticas, ou que incluem palavras que não ocorrem isoladamente, como por exemplo, "ad hoc". Existe ainda um grau maior de divisão dessa classe, de acordo com o tipo de variação permitida:

I. **Expressões Flexíveis Sintaticamente** - possuem uma gama muito grande de variação sintática. Nesta classe encontram-se EMs como:

   A. Construções Verbo-partícula (VPC) - consistem de um verbo combinado com uma ou mais partículas, como por exemplo, *"write up"* e *"look up"*.

   B. Expressões Idiomáticas Decomponíveis - como "*let the cat out of the bag*", tendem a ser sintaticamente flexíveis.

   C. Construções *Light Verbs* - como *"make a mistake"* e *"give a demo"* são altamente idiossincráticas. É difícil prever qual *Light Verb* combina com um determinado substantivo.

II. **Expressões Semifixas** - Algumas EMs são muito rígidas em termos de ordem de palavras e composicionalidade, mas mesmo assim permitem certo nível de variação morfológica *(kick the bucket, kicks/kicking/kicked the bucket).*

III. **Expressões Fixas** - Essa classe é utilizada para descrever expressões que não possuem variação como, por exemplo, *"ad hoc"* e *"in addition"*. Essas expressões não permitem variações morfossintáticas ou modificações internas.

2.2 Técnicas para Extração de Multipalavras

A descoberta automática de tipos específicos de Expressões Multipalavras tem atraído interesse em diversos pesquisadores da área de PLN ao longo dos anos: Baldwin e Villavicencio (2002), em VPCs, Pearce (2002) e Evert e Krenn (2005), em *Collocations* e Nicholson e Baldwin (2006), em Compostos Nominais.

A proposta apresentada por Baldwin e Villavicencio (2002) é uma combinação de métodos para extrair construções verbo-partícula de corpora não anotados que obteve uma performance considerável de 87% de Revocação e de 85% de Precisão, quando avaliada com textos do jornal *Wall Street*. Já Nicholson e Baldwin (2006) investigaram a predição de relações semânticas inerentes para um dado composto nominal usando medidas estatísticas, tais como o "Intervalo de Confiança".

Por outro lado, Zhang et al. (2006) investigaram técnicas para identificação de EMs em geral gerando uma lista de candidatas a EMs semi-automaticamente com técnicas de mineração de erro em analisadores sintáticos e validando-as utilizando uma combinação da World Wide Web como corpus e alguns métodos estatísticos como a Entropia de Permutação. Para cada EM validada era gerado um candidato para uma nova entrada lexical em uma gramática computacional representada pela abordagem *words-with-spaces* (SAG et al., 2002), que consiste em um modelo que considera uma EM como uma entrada lexical única, podendo assim, capturar adequadamente EMs Fixas. Essa adição de entradas lexicais na gramática resultou em um significante aumento da cobertura gramatical. Porém, não foram feitas avaliações da gramática resultante, não sendo possível manipular corretamente todas as EMs utilizando-se dessa abordagem.

Outra abordagem utiliza uma série de técnicas baseadas em métodos estatísticos calculados a partir da frequência em que as palavras ocorrem, para a identificação de

Multipalavras em corpora implementadas em uma ferramenta lexicográfica chamada *Xtract* (SMADJA, 1993).

Ramish, Villavicencio e Boitet (2010) propuseram a ferramenta *mwetoolkit*. O *mwetoolkit* é utilizado para extração automática de Expressões Multipalavras a partir de um corpora monolíngue. Com ele é possível tanto gerar, quanto validar candidatos à EM. A geração é baseada em formas de superfície, enquanto que para a validação, é utilizada uma série de critérios para a remoção de ruído, como algumas medidas de associação, que independem de linguagem como a Informação Mútua, Dice e Máxima Verossimilhança.

Vários outros pesquisadores da área de Processamento de Linguagem Natural já propuseram diversos tipos de técnicas computacionais que tratam da descoberta de termos compostos. Essas técnicas geralmente são acompanhadas por experimentos empíricos que visam a validação dos métodos apresentados.

Com o recente aumento da eficiência e acurácia de técnicas para pré-processamento de textos, como etiquetação morfossintática *(Tagging)* e análise sintática *(Parsing)*, estes podem tornar-se um auxílio extra para o melhor desempenho das técnicas de descoberta de EMs. O processo de *Tagging* consiste na marcação de palavras em um texto com sua respectiva classe morfossintática (verbo, substantivo, adjetivo e outros). Já o *Parsing* corresponde à análise sintática de uma sequência de palavras, respeitando o uso correto da gramática de um idioma. Ambos os processos, classificam as palavras e expressões de um texto gramaticalmente. Estes métodos de pré-processamento permitem manter ou eliminar determinados tipos de expressões e, a partir dessa filtragem, é possível analisar diferentes métodos de descoberta com determinados tipos de expressão.

Neste trabalho serão apresentadas algumas técnicas de extração, sendo que algumas são posteriormente utilizadas no experimento descrito no Capítulo 5. Assim como nos trabalhos citados anteriormente, a maior parte dos métodos são de "Medidas de Associação". Essas medidas são baseadas em fórmulas matemáticas que calculam e interpretam a frequência de dados co-ocorrentes. Para cada par de palavras extraídos de um corpus é calculado um valor relativo da associação entre elas. Muitas medidas de associação são baseadas na hipótese de testes estatísticos, enquanto outras são puramente heurísticas. Por isso, de uma maneira geral, as pontuações de diferentes métodos não podem ser diretamente comparadas. O tratamento posterior para a comparação das medidas é baseado nos *n-best* de cada lista relativa a cada método, levando-se em conta que, em geral, quanto maior o valor da pontuação, maior a probabilidade de um par candidato ser classificado como uma possível EM.

2.2.1 Informação Mútua

A Informação Mútua é uma medida de informação que uma variável contém sobre outra, também conhecido como *Mutual Information* ou apenas IM. Originalmente definida como a informação mútua entre eventos particulares *x* e *y*. Na teoria de probabilidades e da informação, a Informação Mútua é uma medida típica de associação de duas variáveis randômicas. O valor encontrado quantifica uma medida de dependência dessas duas variáveis.

$$MI = \sum_{a,b,c} \frac{n(abc)}{N} log_2 \left[\frac{n(abc)}{n_{\oslash}(abc)} \right] \tag{2.1}$$

Na Fórmula (2.1), *a*, *b* e *c* representam uma palavra cada e *abc* representa um trigrama (três palavras). O *n(abc)* representa a frequência com que o trigrama ocorre no corpus, n_{ϕ} representa a frequência esperada do trigrama e *N* representa o número de palavras em um corpus ou coleção.

2.2.2 Chi-Quadrado

Outra medida típica de associação, o Chi-Quadrado (*Chi-Square* ou χ^2) é um valor de dispersão para duas variáveis de escala nominal, usado em alguns testes estatísticos. Ele calcula a medida que os valores observados se desviam do valor esperado, caso as duas variáveis não estejam correlacionadas. Quanto maior o valor de χ^2 mais significativa é a relação entre a variável dependente e a variável independente.

$$\chi^2 = \sum_{a,b,c} \frac{[n(abc) - n_{\oslash}(abc)]^2}{n_{\oslash}(abc)} \tag{2.2}$$

Na Fórmula (2.2), assim como na medida anterior o termo *n(abc)* representa a frequência que o trigrama ocorre no corpus, $n_{ø}$ representa a frequência esperada do trigrama e N representa o número de palavras no corpus.

2.2.3 Entropia de Permutação

A Entropia de Permutação (*Permutation Entropy* ou EP) (ZHANG et al, 2006) é uma medida de ordem de associação que calcula a importância do ordenamento das palavras que compõem uma EM, quanto mais fixa for a ordem (menos variação no ordenamento), mais chance de ser uma EM. Por exemplo, todas as variações do trigrama *by and large: large and by*, *large by and*, *and large by* e assim por diante são geradas e a distribuição delas é o que indica se o n-grama original é ou não uma EM.

$$PE = -\sum_{(i,j,k)} p(w_i w_j w_k) ln\left[p(w_i w_j w_k)\right]$$
$$p(w_1 w_2 w_3) = \frac{n(w_1 w_2 w_3)}{\sum_{(i,j,k)} n(w_1 w_2 w_3)} \quad (2.3)$$

Na Fórmula (2.3), *w1*, *w2* e *w3* são palavras distintas, onde a soma se dá sobre todas as permutações das palavras e, portanto, sobre todas as posições possíveis das palavras selecionadas. Para isso, as probabilidades (*p*) são estimadas a partir do número de ocorrências de cada permutação de um trigrama.

2.2.4 Dice

O coeficiente Dice é uma medida utilizada frequentemente na Recuperação de Informação que calcula a co-relação entre duas variáveis aleatórias. Na Fórmula (2.4), *fx*, *fy* e *fxy* são as frequências absolutas das palavras *x*, *y* e tanto *x* quanto *y* juntas, respectivamente.

$$Dice(X,Y) = \frac{2 * f_{XY}}{f_X + f_Y} \quad (2.4)$$

2.2.5 Teste-T

Também conhecido como “Teste de Hipótese”, o Teste-T (em inglês, *Student's T-Test*) é um teste estatístico utilizado para determinar se duas amostras poderão ser provenientes de duas populações subjacentes que possuem a mesma média, ou seja, se dois grupos são estatisticamente diferentes um ao outro. Seu cálculo é dado pela Fórmula (2.5).

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{X_1 X_2} * \sqrt{\frac{2}{n}}} \tag{2.5}$$

2.2.6 Outros métodos

Existem ainda, outros métodos capazes de identificar Expressões Multipalavras em textos. Muitos deles foram estudados para o desenvolvimento deste trabalho, entretanto não foram utilizados no experimento, dentre eles: os propostos por Berry-Rogghe e Wulz (1978), Church e Hanks (1990), Kita et al. (1994), Shimohata, Sugio e Nagata (1997), Blaheta e Johnson (2001) e Pearce (2001).

3 RECUPERAÇÃO DE INFORMAÇÃO

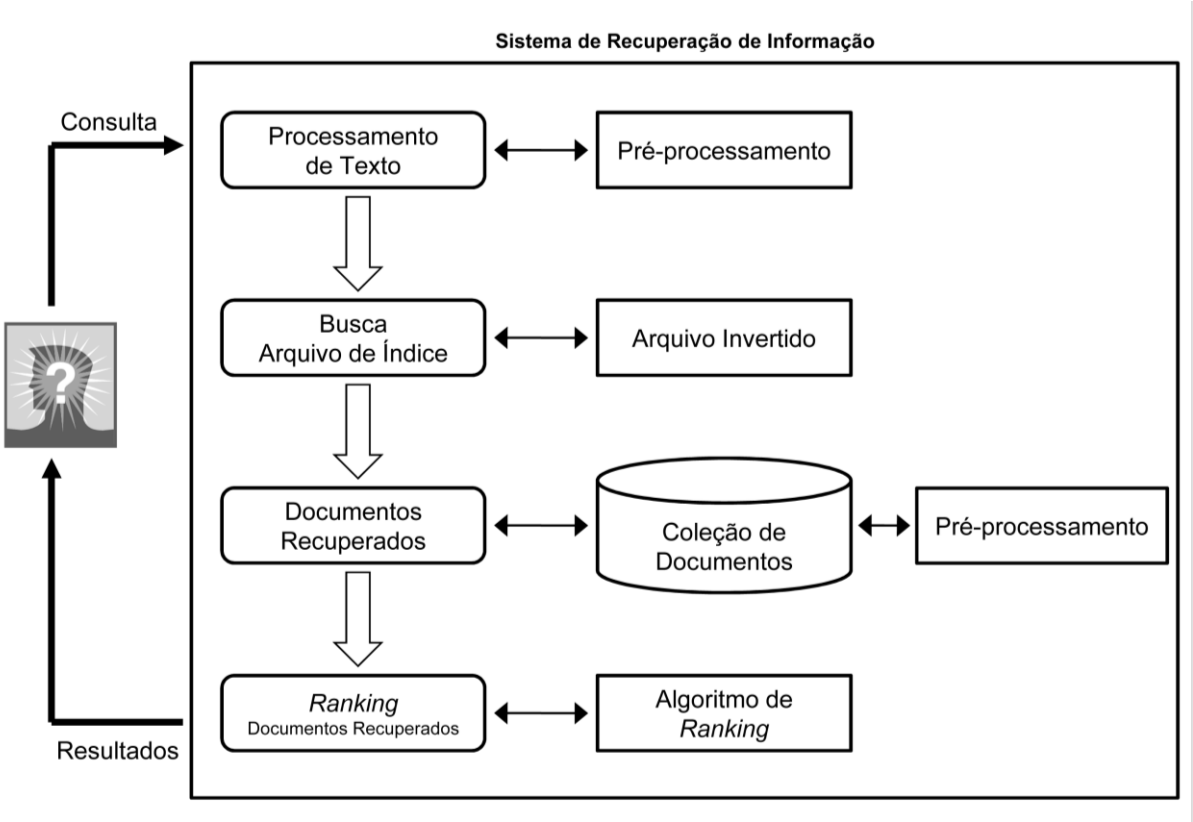
Este capítulo apresenta noções dos conceitos de Recuperação de Informação (RI), também utilizado para o desenvolvimento deste trabalho. Pode-se dizer que desde o surgimento da Ciência da Computação se fazia necessário o estudo de diferentes métodos para o tratamento de informações de forma automática. A RI é a área da computação que basicamente envolve a aplicação de métodos computacionais no armazenamento de documentos para que seja possível recuperar informações associadas a eles de forma automática. Segundo, Manning e Shütze (1999), a Recuperação de Informação é definida por:

> "Encontrar material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação nas grandes coleções (usualmente armazenadas em servidores locais ou na internet)."

Baseados nesta definição foram criados diferentes tipos de Sistemas de Recuperação de Informação (SRI). Um SRI pode ser definido como um conjunto de dados padronizados, armazenados em meio eletrônico, utilizados para identificar informação e fornecer sua localização (Figura 3.1). Assim, pode-se dizer que o principal objetivo de um SRI é recuperar informação contida em documentos, que possam ser úteis ou relevantes para um usuário, ou seja, a "necessidade de informação do usuário". Tais documentos geralmente são arquivos de texto, porém existem outros métodos que tornam possível a recuperação de informação em outros tipos de mídias como sons, vídeos, imagens, entre outros. Para que um usuário satisfaça sua "necessidade", ele precisa traduzir a informação em uma "consulta". Esta consulta geralmente é representada através de um conjunto de palavras-chave no mecanismo de busca do sistema de RI. As palavras-chave utilizadas são consultadas de forma automatizada nos dados padronizados e retornam os documentos estimados como de interesse do usuário através de um *ranking*. Uma possível inconveniência imediata dessa abordagem é que o uso de palavras-chave usualmente introduz uma diferença de semântica entre a intenção do usuário e o conjunto de documentos retornados, por exemplo, a palavra "posto" pode ser usada como: "Ele ocupa um alto posto na empresa. / Abasteci meu carro no posto da esquina". Além disso, essa diferença de semântica pode ser ampliada devido à dificuldade adicional em se lidar com textos em linguagem natural, que nem sempre são bem estruturados e podem ser semanticamente ambíguos, como por exemplo, "*The food is hot*" não é possível definir se trata-se da temperatura da comida ou se ela é apimentada. Como resultado, grandes

chances de presença de documentos não relevantes entre os documentos retornados por uma consulta. Nesse cenário, o principal objetivo dos SRI foca-se em recuperar o maior número possível de documentos relevantes e o menor número possível de documentos não relevantes (WIVES; OLIVEIRA, 2001).

Figura 3.1: Modelo Básico - SRI



Fonte: Adaptado de Orengo, 2004.

Na maioria dos Sistemas de Recuperação de Informação, o conjunto de documentos a ser utilizado necessita passar por algum tipo de pré-processamento, onde seus conteúdos são decompostos para que fiquem disponíveis para serem recuperados. Este pré-processamento geralmente envolve, principalmente:

- **Tokenização** - refere-se à decomposição dos documentos em cada termo ou palavra que os compõem. Em alguns casos, é difícil definir o que se entende por uma "palavra", mas geralmente os delimitadores utilizados para a separação dos termos são os espaços em branco, tabulações, quebras de linhas ou outros caracteres especiais. No entanto, em algumas línguas como a Chinesa, o processo de tokenização é mais complexo, pois não existem limites para separar uma palavra.

- ***Stemming*** - consiste no processo de redução das formas variantes das palavras em apenas um radical, geralmente pela remoção de seus sufixos. Por exemplo, as palavras da língua inglesa: *presentation*, *presented*, e *presenting* pode ser reduzido para o radical *present*. Isso é baseado na suposição de que uma consulta com o termo *presenting* implica no interesse em documentos que contêm as palavras *presentation* e *presented*. Com a utilização deste processo, os termos derivados de um mesmo radical são contabilizados como um único termo. Assim, o número de termos a serem indexados é reduzido, mas o número de documentos por termo é aumentado, visando assim aumentar a revocação. Este processo ainda diminui o tamanho do índice, tornando mais ágil a recuperação nos sistemas.

- **Remoção de *Stopwords*** - *stopwords* são palavras comuns no uso de uma língua, consideradas irrelevantes para a consulta/indexação de documentos. Geralmente, são preposições, conjunções, pronomes e artigos. Uma palavra também pode ser considerada uma *stopword* caso ela apareça um considerável número de vezes dentro de um texto (parâmetro configurável). Assim, esta palavra não representa um significado relevante para a recuperação e pode influenciar diretamente nos resultados, podendo então ser descartada. Além de eliminar palavras de baixa significância, a remoção das *stopwords* também pode ser usada na redução de textos, podendo diminuir seu tamanho em até 50%, aumentando a performance do sistema e diminuindo o espaço ocupado em disco.

- **Cálculo de Frequência** - a contagem de termos é utilizada para medir a importância dos termos que compõem os documentos. A ideia da atribuição de pesos para as palavras, basicamente resume-se em atribuir pesos baixos para palavras que têm uma frequência alta em documentos e valores de peso maiores para palavras que aparecem poucas vezes. Existem diferentes métodos automáticos para se atribuir peso para termos, porém um dos mais conhecidos é o **tf-idf** (*term frequency-inverse document frequency*), que consiste em uma função entre o número de ocorrências de um termo em um documento e do número de ocorrências deste termo em toda a coleção de documentos. Este método é descrito pela Fórmula (3.1).

$$w_{ij} = tf_{ij} \times \log_2 \frac{N}{n} \quad (3.1)$$

Onde w_{ij} representa o peso atribuído ao termo T_j no documento D_i e tf_{ij} representa a frequência do termo *Tj* no documento *Di*. O número de documentos da coleção é representado por *N* e o número de documentos em que o termo T_j aparece é representado por *n*.

- **Indexação** - envolve a criação de estruturas de dados internas, associadas aos documentos de uma coleção, que permitam consultas de forma rápida em textos. Uma das estruturas mais utilizadas são os "arquivos invertidos". Os arquivos invertidos possuem um dicionário, que consiste em uma lista de todos os termos da coleção que foram armazenados na base de dados e uma lista de documentos, com as posições de onde cada termo ocorre. Este tipo de estrutura é eficiente em termos de acesso, entretanto seu consumo em disco varia entre 10% ou mais de 100% do tamanho original do documento.

3.1 Modelos Clássicos de Recuperação de Informação

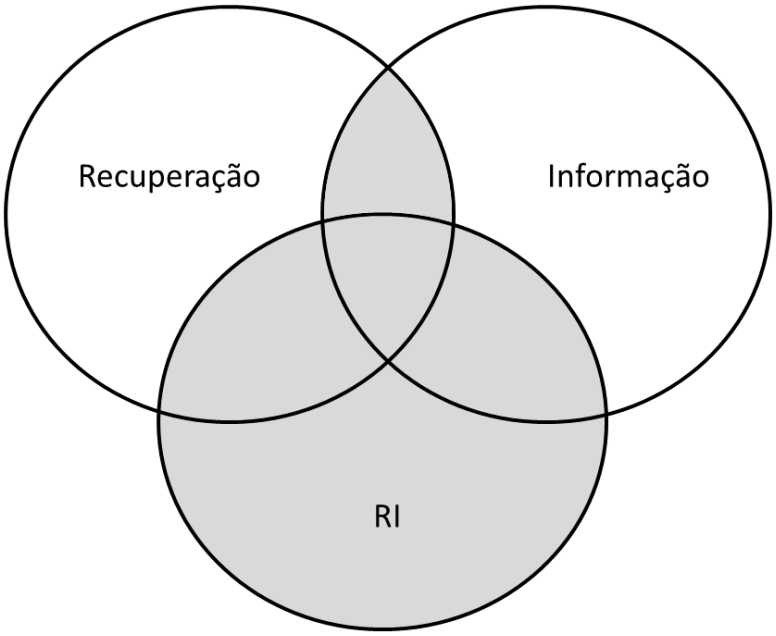
Para gerar uma classificação ordenada dos resultados, um sistema de RI usualmente adota um modelo para representar os documentos e as consultas do usuário. Muitos modelos ou abordagens para o cálculo da classificação têm sido propostos ao longo dos anos, sendo três modelos considerados clássicos: Modelo Booleano (Subseção 3.1.1), Modelo Vetorial (Subseção 3.1.2) e o Modelo Probabilístico (Subseção 3.1.3). Estes modelos serão apresentados nas subseções a seguir.

3.1.1 Modelo Booleano

O modelo booleano é um dos modelos clássicos que considera uma consulta como uma expressão booleana convencional, que liga seus termos através dos conectivos lógicos *AND* (união), *OR* (intersecção) e *NOT* (exclusão ou negação). No modelo booleano, um documento é considerado relevante ou não relevante a uma consulta, não existe resultado

parcial e não há informação que permita a ordenação do resultado da consulta. Por exemplo, a informação a ser recuperada pela consulta (“*Recuperação*” **AND** *“Informação”*) **OR** *“RI”* consequentemente resultaria em todos os documentos que contenham as palavras *Recuperação* e *Informação* no mesmo documento, ou apenas contenha a abreviação *RI*, como é mostrado na Figura 3.2.

Figura 3.2: Exemplo – Modelo Booleano



Fonte: Próprio autor.

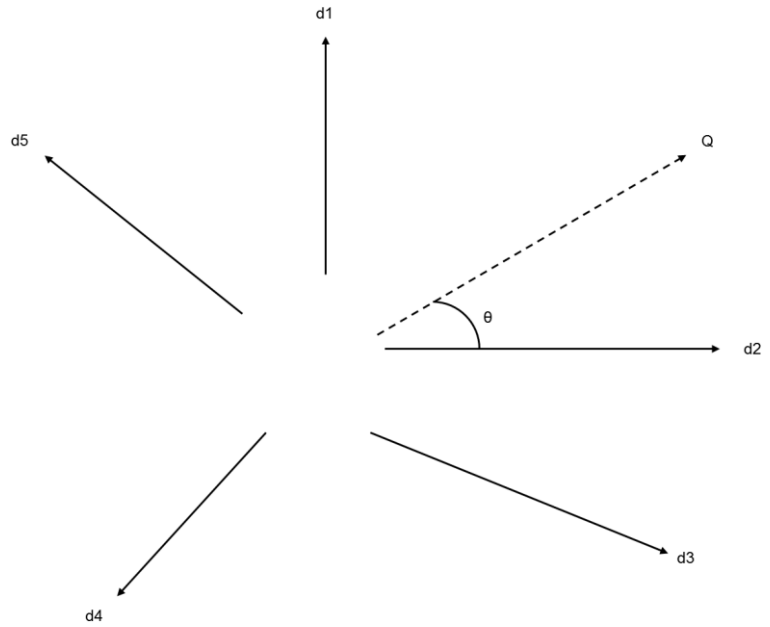
Este modelo ainda é muito utilizado, principalmente em sistemas comerciais, apesar do inconveniente de se classificar apenas como relevante ou não relevante, sem a possibilidade de ranqueamento do resultado. Um outro problema do modelo booleano é de ele não ser capaz de identificar a importância de um termo em um documento, ou seja, se um termo está presente em um documento ele sempre é considerado muito importante, não havendo valores intermediários. Entretanto, houve propostas de extensão deste modelo, com suporte a cálculo de frequência para que fosse possível dar um peso aos termos e posteriormente apresentá-los em um *ranking*, porém não são muito utilizadas (SALTON; FOX; WU, 1983).

3.1.2 Modelo Vetorial

Proposto por Salton (1971), este modelo atribui pesos para os termos contidos nos documentos e nas consultas. Cada documento é representado por um vetor de termos onde cada termo possui um valor associado, indicando seu grau de importância no documento. Assim, cada documento possui um vetor associado, constituído por pares de elementos: *(termo_1, peso_1), (termo_2, peso_2)...(termo_n, peso_n)*. Os termos que não existem no documento, recebem o peso 0 (zero) e os termos que constam nos documentos são calculados através de uma fórmula que identifica seu grau de importância, fazendo com que pesos próximos a 1 (um) sejam termos muito relevantes (dependendo da normalização, a faixa de peso pode variar de -1 a 1).

Nos vetores são representadas todas as palavras da coleção e não somente as presentes no documento. Cada elemento do vetor é considerado uma coordenada dimensional e podem ser colocados em um espaço euclidiano de *n* dimensões (sendo *n* o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso. O grau de similaridade entre os documentos é medido pelo ângulo entre esses vetores, assim como a similaridade com a consulta do usuário, também representada por um vetor. Assim, os vetores que estiverem próximos, em uma mesma região, em teoria tratam de um assunto similar. A Figura 3.3 exemplifica esse modelo, onde *d* são vetores de documentos e *Q* é o vetor de consulta.

Figura 3.3: Exemplo – Modelo Vetorial



Fonte: Próprio autor.

O cálculo para atribuir um peso ao termo, pode ser calculado de diversas maneiras, porém umas das formas mais utilizadas é a função Cosseno (*Cosine*) (SALTON; BUCKLEY, 1987), que calcula o produto dos vetores através da Fórmula (3.2).

$$sim(Q,D) = \frac{\sum_{k=1}^{n} w_{qk} * w_{dk}}{\sqrt{\sum_{k=1}^{n} (w_{qk})^2} * \sqrt{\sum_{k=1}^{n} (w_{dk})^2}} \quad (3.2)$$

Nessa fórmula, *Q* representa o vetor de termos da consulta, *D* representa o vetor de termos do documento, w_{qk} são os pesos dos termos da consulta e w_{dk} são os pesos dos termos dos documentos. No caso da função *Cosine*, os documentos mais relevantes são os que têm o cosseno do ângulo dos vetores de documentos e de consulta próximos à 1 (um), por outro lado os mais irrelevantes são os documentos em que o cosseno do ângulo seja próximo à -1 (menos um).

Por fim, depois de calculados os graus de similaridade, é possível montar um *ranking* de todos os documentos com seus respectivos graus de relevância à consulta. Sendo ordenada da mais alta a mais baixa similaridade, para que o usuário identifique os documentos relevantes.

3.1.3 Modelo Probabilístico

Consiste em um conjunto de modelos considerados probabilísticos que utilizam conceitos provenientes da área de probabilidade e estatística. Proposto por Robertson e Sparck-Jones 1976, ele também é conhecido como *Binary Independence Retrieval*. Nele, o objetivo é identificar a probabilidade de um documento *x* ser relevante em uma consulta *y*, caso existam termos em comum entre eles. Entretanto, o modelo probabilístico é baseado no conceito de um termo estar ou não presente em um documento e, para isso, é atribuído a cada documento recuperado, um valor de relevância para uma determinada consulta. O modelo probabilístico pressupõe a independência dos termos indexados, o que representa que a relevância de um documento não tem nenhuma relação com a relevância de outro documento.

Basicamente, dada uma consulta de um usuário, existe um conjunto que contém exatamente os documentos relevantes e este é o resultado ideal a ser recuperado. Dada uma consulta e um documento, o modelo probabilístico tenta estimar a probabilidade em que o

documento é relevante para a consulta em questão. Os documentos são ranqueados em ordem decrescente, de acordo com a relevância calculada. Desta forma, o modelo assume que a probabilidade calculada para determinar a relevância depende apenas da representação da consulta e do documento. A similaridade de uma consulta e um documento é dada pela razão entre a probabilidade de um documento ser relevante para a consulta e a probabilidade do documento não ser relevante para a mesma, como pode ser observado pela Fórmula (3.3), onde *Q* representa a consulta e *d* representa um documento.

$$sim(d_i, Q) = \frac{P(relevante \mid d_i)}{P(nao\_relevante \mid d_i)} \tag{3.3}$$

No início do processo de recuperação em um modelo probabilístico, não há informações sobre a relevância dos documentos, portanto uma estimativa de probabilidade inicial é feita com base na distribuição das palavras nos documentos. Uma hipótese seria que a probabilidade de um documento ser relevante fosse constante para todos os termos de índice, como por exemplo: 0,5 (Fórmula (3.4)). E a probabilidade dos termos entre os documentos não relevantes pode ser medida através da distribuição desses termos em toda a coleção (Fórmula (3.5)).

$$P(k_i \mid relevante) = 0,5 \tag{3.4}$$

$$P(k_i \mid nao\_relevante) = \frac{n_i}{N} \tag{3.5}$$

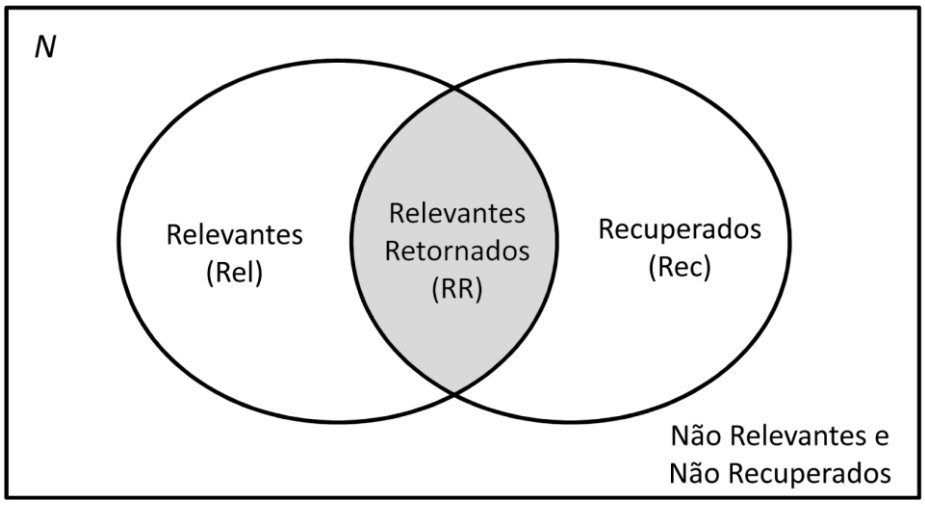
Onde $P(k_i \mid relevante)$ representa a probabilidade de um termo k_i estar presente em um documento relevante, $P(k_i \mid nao\_relevante)$ representa a probabilidade de um termo k_i estar presente em um documento não relevante, n_i corresponde ao número de documentos em o termo aparece e *N* corresponde ao número de documentos da coleção.

Após obter a recuperação do conjunto inicial, é possível que o cálculo de probabilidade de relevância possa ser melhorado, a partir da repetição recursiva do processo, julgando se os itens são relevantes ou não.

3.2 Métricas de Avaliação em Recuperação de Informação

Para que seja possível avaliar os resultados obtidos por um SRI é necessária a utilização de métricas que possibilitem obter a informação se a consulta do usuário funcionou como deveria, ou seja, de acordo com a sua necessidade. As métricas podem informar quais e quantos documentos são relevantes, além do quanto cada um deles é relevante. Para o cálculo correto dessas métricas é necessário que os documentos utilizados pelo SRI sejam previamente marcados para quais consultas são relevantes. A eficiência e eficácia de um SRI é avaliada de acordo com a sua capacidade em recuperar o máximo possível de itens relevantes ao mesmo tempo em que elimina o maior número possível de documentos não relevantes (Figura 3.4).

Figura 3.4: Resultado de uma Consulta em um SRI



Fonte: Próprio autor.

Na computação, existem diversas métricas que são utilizadas para a avaliação de resultados e desempenho de um SRI, a seguir serão apresentadas duas das mais comuns para a área de RI.

- **Precisão** (Precision) - esta métrica mede a habilidade do SRI em retornar apenas documentos relevantes, o que evita que o usuário desperdice tempo com documentos não relevantes que foram recuperados na consulta (Fórmula (3.6)).

$$P = \frac{RR}{Rec} \tag{3.6}$$

Onde *RR* são os documentos relevantes retornados e *Rec* são os documentos recuperados.

- **Revocação** (Recall) - esta métrica mede a habilidade do SRI em retornar o maior número de documentos relevantes, entretanto requer o conhecimento, a priori, de quantos documentos relevantes existem na coleção (Fórmula (3.7)).

$$R = \frac{RR}{Rel} \tag{3.7}$$

Onde *RR* são os documentos relevantes retornados e *Rel* são os documentos relevantes.

Como exemplo, suponhamos que 20 documentos são retornados de uma consulta, 7 deles são relevantes e o número total de documentos relevantes para essa consulta é 10. Dessa, forma a revocação será de 70% (R = 7/10) e a precisão de 35% (P = 7/20). Usualmente, altos valores de revocação contrastam com baixos valores de precisão e vice-versa.

Outras métricas de avaliação que posteriormente serão usadas por este trabalho são:

- **Precisão Média** *(Average Precision - AvP)* - é a média das precisões computadas após cada documento relevante recuperado, buscando medir o percentual de documentos relevantes presentes no topo do *ranking*.
- **Média da Precisão Média** *(Mean Average Precision)* - conhecida como *MAP*, esta métrica é uma medida padrão para a análise de resultados ranqueados. Portanto, o *MAP* é utilizado para conjuntos de consultas, onde o valor é dado a partir da média dos escores de precisão média *(AvP)* para cada consulta (Fórmula (3.8)).

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} P(R_{jk}) \tag{3.8}$$

Onde /Q/ é o número de consultas, R_{jk} consiste no conjunto de resultados ranqueados a partir do melhor resultado até o documento d_k e m_j é o número de documentos relevantes para a consulta j.

- **Precisão-R** *(R-Precision*) - é a precisão até a posição R do *ranking*. Onde R é o número total de documentos relevantes. A Precisão-R consiste no cálculo da precisão no determinado ponto em que a quantidade de documentos recuperados se iguala ao número de documentos relevantes a serem recuperados.

Várias iniciativas para avaliações uniformes de SRIs têm sido propostas, como por exemplo, o *Cross-Language Evaluation Forum* (CLEF). O CLEF é uma forma imparcial de avaliação que ocorre em várias etapas, durante um período de oito (8) meses. Os participantes recebem as coleções de teste e têm cerca de 3 meses para trabalharem com elas. No estágio seguinte, os participantes recebem os tópicos das consultas e enviam os resultados à coordenação do CLEF em um arquivo texto com uma lista, contendo os documentos recuperados para cada tópico em ordem decrescente de relevância. É então feita a comparação de sistemas diferentes com base em seu desempenho. A campanha finaliza com um workshop reunindo os participantes para apresentação de suas abordagens (CLEF, 2008).

3.3 Expressões Multipalavras e a Recuperação de Informação

O foco deste trabalho é a investigação de possíveis melhorias na indexação de Expressões Multipalavras em sistemas de Recuperação de Informação. A integração de métodos de Processamento de Linguagem Natural é bastante utilizada em sistemas de RI. Como exemplo, pode-se citar o *Stemming* que como já foi visto, consiste no processo de redução das formas variantes das palavras em apenas um radical. Muitas vezes o radical encontrado pelo processo de *Stemming* pode ser diferente do radical morfológico da palavra. O *Stemming* é muito útil em mecanismos de busca para uma expansão de consulta ou indexação e outros problemas de Processamento de Linguagem Natural. Estima-se uma melhoria de 1% a 45% em sua utilização, tendo seu desempenho ainda aumentado em coleções de documentos pequenos. Outro tipo de pré-processamento de texto utilizado é a remoção de *Stopwords*. As *stopwords* são palavras consideradas irrelevantes para a análise de um determinado tipo de texto. Entretanto, apesar do processo de remoção de *stopwords* ser

amplamente usado para eliminar "ruídos" em textos, isso não significa que ele acarrete em melhorias na indexação, principalmente se posteriormente houver tentativas de extração de EMs. Por exemplo, a expressão "Banco de Dados", pode significar uma área da Computação ou um conjunto de registros dispostos em uma estrutura regular que possibilita a organização dos mesmos. Em um sistema comum de RI a preposição *de* seria removida pelo processo de remoção de *stopwords* e seriam indexadas apenas as palavras "banco" e "dados" separadamente. Cada uma dessas palavras possui significados diferentes dos usados no termo "Banco de Dados", portanto a remoção de *stopwords* é prejudicial, pois seria perdido o verdadeiro significado proposto no texto. Por isso, sugere-se a indexação de Multipalavras nesse tipo de sistema para que o significado real do termo seja preservado.

Para que fosse validada a hipótese que EMs são benéficas para a RI, foram executados experimentos sobre este assunto. Estes foram submetidos[1] ao *Cross-Language Evaluation Forum (CLEF)* (ACOSTA et al., 2008) e comprovaram melhorias na utilização de EMs em mecanismos de RI. Neste caso, a identificação das Expressões Multipalavras utilizou-se de medidas de associação como a Informação Mútua (Seção 2.2.1) e o Chi-quadrado (Seção 2.2.2). Para o cálculo das medidas foi necessário extrair a frequência de cada termo que compõe um bigrama (duas palavras) e também do próprio bigrama em si. Assim, dois *rankings* foram criados, sendo um de cada medida. Para que fosse possível identificar as melhores candidatas a EMs, esses *rankings* foram unificados, formando apenas um, de onde foram retiradas as melhores n candidatas a EMs. As melhores 7.500 candidatas foram adicionadas em um SRI como unidades únicas em um corpora jornalístico, cedido pelo CLEF, assim como tópicos de consulta para treinamento e teste. Foram elaboradas diferentes variações de como proceder o experimento, mesclando a utilização de EMs como termos únicos e a utilização ou não de *Stemmers*, sempre comparados a um *baseline* (texto original). Primeiramente, com os tópicos de consulta de treinamento foi atingida uma grande melhora na recuperação de documentos relevantes com a utilização de Expressões Multipalavras. Entretanto, os resultados atingidos pelos tópicos de testes não mostraram uma melhora significativa, restringindo-se apenas à melhora de alguns tópicos de consulta. Com esses resultados, este experimento atingiu a quinta posição na classificação da *Robust-WSD Task* do CLEF 2008, encorajando assim a continuação dos estudos na utilização de EMs em SRI.

[1] Acosta, O.C., Geraldo, A.P., Orengo, V.M., Villavicencio, A.: UFRGS@CLEF 2008: Indexing Multiword Expressions for Information Retrieval. In: Borri, F., Nardi, A., Peters, C. (eds.) Working Notes for the CLEF 2008 Workshop (2008), http://www.clef-campaign.org/

Ainda existem muitos sistemas de RI que resistem à utilização de métodos de PLN, pois os consideram lentos e com uma eficiência muito fraca. Porém, com os avanços nos métodos e recursos da PLN, esse cenário está mudando e acredita-se que a integração dessas áreas traga grandes benefícios à comunidade científica.

4 MATERIAIS E MÉTODOS

Baseado na hipótese de que as Expressões Multipalavras possam aprimorar o resultado de sistemas de Recuperação de Informação, neste capítulo serão apresentados os diferentes recursos e métodos utilizados para a realização do experimento deste trabalho. Este, será descrito posteriormente, no Capítulo 5, de forma mais detalhada, bem como os objetivos específicos para cada tipo de avaliação.

4.1 Recursos e Ferramentas

Para que fosse possível avaliar a utilização de Expressões Multipalavras em sistemas de Recuperação de Informação foi necessária a realização de experimentos com o objetivo de verificar diferenças em performance, com ou sem identificação de EMs nos resultados obtidos por sistemas de RI. Esse tipo de avaliação requer o uso de um corpus consistente e de tamanho considerável, contendo uma alta diversidade de termos. Para isso, foram utilizados os corpora jornalísticos:

- Los Angeles Times (Los Angeles, EUA – 1994)
- The Herald (Glasgow, Escócia – 1995)

Ambos os corpora abrangem diversos tipos de notícias divulgadas por essas mídias nos anos citados. A língua utilizada é o inglês americano, no caso do Los Angeles Times e o inglês britânico, no caso do The Herald. Por simplificação, o corpus do Los Angeles Times será referido neste trabalho como LA94 e o The Herald como GH95. Ambos possuem pouco mais de 160.000 notícias (Tabela 4.1), sendo cada notícia considerada um documento.

Tabela 4.1: Total de Documentos

Corpus	Documentos	Total de Termos	Termos Distintos	Tamanho (MB)
LA94	110.245	59.978.338	362.495	348
GH95	56.472	23.614.046	194.398	136
Total	166.717	83.592.384	469.785	484

Fonte: Próprio autor.

A coleção de documentos, assim como os tópicos de consulta e a relação de documentos relevantes, que serão comentados posteriormente, foram disponibilizados pelo CLEF 2008 (*Cross Language Evaluation Forum*), na tarefa intitulada *Robust-WSD*. Essa tarefa visava explorar a contribuição da desambiguação das palavras na RI monolíngue ou

bilíngue. A desambiguação de sentido refere-se a determinação do significado de um termo dado um contexto, para termos que possuem mais de um sentido. Dessa forma, os documentos do corpus foram anotados por um sistema de desambiguação. Um exemplo da estrutura dos documentos é apresentado na Figura 4.1.

Figura 4.1: Estrutura dos Documentos Originais

```
<TERM ID="GH950102-000000-126" LEMA="underworld" POS="NN">
<WF>underworld</WF>
<SYNSET SCORE="0.5" CODE="06120171-n"/>
<SYNSET SCORE="0.5" CODE="06327598-n"/>
</TERM>
```

Fonte: Próprio autor.

A estrutura de um documento contém informações sobre o identificador de um termo em um documento (*TERM ID*), o lema de um termo (*LEMA*) e também sua etiqueta morfossintática (*POS*). Ainda possui a forma em que o termo aparecia no texto (*WF*) e informações do termo na WordNet (MILLER, 1995) (FELLBAUM, 1998) como *SYNSET SCORE* e *CODE*. Neste trabalho, foram extraídos os termos localizados no campo *"LEMA"*, que como o próprio nome diz, consiste no lema de uma palavra, ou seja, sua forma canônica. A utilização dos lemas e não das palavras para a formatação do corpus evita que variações linguísticas comprometam os resultados dos experimentos. Por exemplo, as variações do verbo *to write* como *wrote* e *written* são mantidas como *"to write"*, o que facilita a busca por documentos que, em teoria, tratam do mesmo assunto.

Desta forma, os documentos do corpus foram formados apenas por lemas e o passo seguinte é a indexação destes documentos, com o uso de um sistema de RI. Para esta tarefa utilizou-se uma ferramenta chamada *Zettair* (ZETTAIR, 2008), que consiste em um mecanismo de busca textual compacto e ágil, podendo ser utilizado tanto para a indexação quanto para busca de informações em coleções de texto. Uma de suas principais características é a habilidade de manipular grandes coleções de documentos, além de ser uma ferramenta de código aberto, desenvolvida por um grupo de pesquisadores da *The Royal Melbourne Institute of Technology University*. A indexação é uma etapa simples do processo, pois apenas apontam-se os arquivos que contêm os documentos que compõem o corpus e o próprio sistema de RI os organiza, para que possam ser executadas buscas de documentos de forma ágil. A única opção utilizada na indexação é o *stemmer*, que no caso do *Zettair* utiliza o

Porter Stemming (PORTER, 1997) por padrão e este foi o adotado para este trabalho. O *stemming* pode proporcionar um maior vínculo entre termos relacionados. Por exemplo, *bomb* e *bombing* não eram associados ao mesmo termo nos textos lematizados, entretanto após o *stemming*, foram representados de forma única, constituindo um único termo na estrutura do índice.

Após a indexação, a próxima etapa consistiu na preparação dos tópicos de consulta. Da mesma forma que o corpus, apenas o lema dos tópicos foram extraídos e utilizados para formar um documento contendo 310 tópicos de consulta. Estes tópicos são numerados de 41 a 350 e resumem-se em pequenos fragmentos de texto utilizados para retornar um ou vários documentos considerados relevantes para determinado tópico. A determinação de que um documento é ou não relevante para cada tópico foi dada de acordo com a relação de documentos relevantes, manualmente preparada e fornecida juntamente com o material disponibilizado pela organização do CLEF. Com o arquivo de tópicos criado, é utilizado um utilitário do *Zettair*, chamado *zet_trec*. Este é responsável pela geração dos resultados na busca por documentos para cada um dos tópicos. Para este trabalho, foi configurada uma saída de 1.000 resultados por tópico, ou seja, os primeiros 1.000 documentos que obtiveram maior escore na consulta de determinado tópico. Outra configuração foi o uso da métrica do cosseno para o cálculo do escore, que é a medida padrão do método vetorial.

Por fim, após a indexação e geração do escore na busca de documentos, é verificada a quantidade de relevantes retornados. Para isso, utiliza-se a ferramenta *trec_eval*, disponibilizada também pela comunidade do CLEF. Com ela é possível fazer uma comparação entre os resultados obtidos com *zet_trec* e a relação de documentos relevantes. Com isso, teremos os resultados finais para cada "rodada" de experimentos, sendo relatado, entre outros, a quantidade total de relevantes retornados, a precisão média e a precisão-R, como os mais importantes para a avaliação dos resultados.

4.2 Inserção de Expressões Multipalavras como Termo Único

Para verificar o impacto do uso de Expressões Multipalavras em um SRI, algumas das rodadas do experimento tiveram o acréscimo de Expressões Multipalavras nos documentos do corpus e nos tópicos. Essas expressões estavam contidas no próprio texto e foram duplicadas no intuito de forçar o sistema de RI a considerar EMs na recuperação de informações relevantes, específicas dessas expressões. No caso deste trabalho, foram utilizadas apenas

expressões compostas por duas palavras, ou seja, somente bigramas. Cada bigrama (Expressão Multipalavra) contido em um dicionário pré-definido e que ocorra em determinado documento, é tratado como um termo único, para fins de indexação.

Espera-se que desta forma, os documentos que contenham determinada EM sejam indexados com maior precisão do que os que contêm as mesmas palavras da expressão separadamente, gerando assim um percentual de acerto maior.

4.3 Dicionários de Expressões Multipalavras

Neste trabalho, foram gerados dicionários contendo Expressões Multipalavras que posteriormente são inseridas no corpus, como um termo único. Estes dicionários são criados a partir de técnicas de extração automáticas e manuais. Os diferentes métodos de extração estão relacionados com a hipótese de que uma melhor qualidade das EMs presentes nos dicionários podem trazer melhores benefícios para um sistema de RI. Em outras palavras, a escolha de melhores candidatos à EM podem acarretar em melhores resultados na recuperação. Com alternativas automáticas para a criação de dicionários, é possível evitar métodos manuais custosos e que dependem de um especialista. Entretanto, métodos automáticos estão sujeitos a conterem ruídos, selecionando falsas candidatas. Os objetivos da criação destes dicionários ficam então voltados na comparação entre as diferentes possibilidades de se classificar EMs positivas. Para isso foram utilizados quatro diferentes métodos para a geração de sete dicionários de EMs, conforme descritos nas subseções 4.3.1, 4.3.2, 4.3.3, 4.3.4.

4.3.1 Métodos Estatísticos

Para a criação deste dicionário foram extraídos todos os bigramas contidos no corpus. Como a quantidade de bigramas existentes era muito grande (99.744.811 bigramas) utilizou-se então, informações que existiam nos documentos originais, as etiquetas morfossintáticas. Junto com o campo *"LEMA"*, extraído para o procedimento anterior, também foi extraído o campo "POS" (*part-of-speech*), que descreve a classe gramatical de cada palavra. Foi definido então que, para tornar mais viável e ágil o experimento, serão utilizados apenas bigramas formados por Compostos Nominais, ou seja, quando o campo *POS* estivesse preenchido com o valor NN (*Noun*). Desta forma, com bigramas compostos por sequências de NN e um pré-

processamento para eliminar ruídos que pudessem prejudicar o experimento, a quantidade de bigramas candidatas à EMs foi então reduzida para 308.871 bigramas.

1. **Compostos Nominais -** Com essa listagem de compostos nominais candidatos à EM, o próximo passo efetuado foi a seleção de bigramas que tivessem uma frequência maior no texto, assim foram escolhidas as EMs que ocorrem 10 vezes ou mais em todo o corpus. Isso foi proposto pelo fato de não favorecer termos que ocorram poucas vezes em um determinado domínio, o que pode facilitar a busca por documentos relevantes e também como um possível limiar para considerar expressões da coleção como Expressões Multipalavras. Assim, foi criada a primeira lista de EMs, composta por 15.001 bigramas, a qual será chamada de dicionário **D1**.

2. **Melhores Compostos Nominais -** Após a criação do primeiro dicionário, foi possível refiná-lo com a utilização de métodos estatísticos. Neste caso, os métodos utilizados foram a Informação Mútua e o Chi-Quadrado (conforme ACOSTA et al. (2008)). Para este cálculo é necessário um valor que corresponda à frequência de utilização da EM, bem como a utilização individual de cada palavra na língua inglesa. Desta forma foi necessário obter esses valores de frequência pela WEB, utilizando a ferramenta de busca do *"Yahoo!"*, pois apesar da grande quantidade de termos no corpus, podia ocorrer o fato de que determinada EM ocorresse em grande quantidade no domínio jornalístico. Com os dados necessários para o cálculo, foi gerado um ranking com o escore decrescente para cada método estatístico. Para que fosse possível juntar os *rankings*, foi calculada a média entre as posições de cada EM e as 7.500 primeiras colocadas formam o segundo dicionário de EMs e será chamado de **D2**. Por exemplo, uma candidata à EM foi ranqueada na segunda posição no método IM e na quarta posição no método χ^2. A média entre os dois valores de posição resultaria em "3", dessa forma, as 7.500 candidatas que obtiveram o menor valor, foram selecionadas para compor esse dicionário.

3. **Compostos Nominais Menos Frequentes -** Este dicionário foi gerado a partir de bigramas de compostos nominais que tinham frequência no corpus inferior a 10 e superior a quatro ocorrências. Ele foi criado com o intuito de avaliar se a escolha de EMs com baixa frequência, consequentemente acarretasse em uma

piora nos resultados, comparado principalmente ao dicionário anterior. Estas EMs formam o terceiro dicionário, com 17.328 bigramas, chamado de **D3**.

4.3.2 Gold Standard

Este dicionário foi criado a partir de uma sublista do CIDE (CAMBRIDGE, 1995), no intuito de avaliar como a qualidade máxima de EMs presentes em dicionários manualmente criados e de ampla cobertura, possam afetar os resultados. O dicionário de expressões do CIDE consiste em uma série de termos compostos previamente estudados e classificados como Expressões Multipalavras verdadeiras. Como essa lista abrange todos os tipos de EMs, foi necessário a utilização do TreeTagger (SCHMID, 1994), que consiste em uma ferramenta para a etiquetação morfossintática de palavras, com eficiência de 96% para a língua inglesa, e para o presente trabalho foi utilizada em sua configuração padrão. Esta ferramenta foi utilizada para fossem extraídos apenas os compostos nominais do dicionário do CIDE, ou seja, apenas as EMs que fossem formadas por sequências de NN. Após foram selecionados apenas as EMs que também existiam no corpus, para que assim fossem inseridas nos locais em que o termo ocorria. Desta forma, 568 EMs formaram o quarto dicionário e ele será chamado de **D4**.

4.3.3 Árvore de Decisão

Outra abordagem para a criação de um dicionário de Expressões Multipalavras foi o uso de um classificador para combinar medidas estatísticas. Cada medida calcula de maneira peculiar a relação entre termos e isso justifica o motivo em combiná-las para uma classificação. Por exemplo, a Informação Mútua indica a dependência entre duas palavras, por outro lado a Entropia de Permutação indica a importância na ordem em que elas ocorrem no texto. O classificador utilizado para a criação deste dicionário é uma árvore de decisão[2]. Uma árvore de decisão (QUINLAN, 1986) pode ser definida como uma estrutura em forma de árvore que, dada uma série de valores de parâmetros de um objeto, o classifica em uma determinada classe. Para isso é necessário fornecer dados de treinamento com candidatas à EMs verdadeiras e falsas em quantidade equilibrada, para que o resultado não favoreça

[2] Neste trabalho foi utilizado o algoritmo J48 (WITTEN; FRANK, 2000) que é uma implementação do algoritmo C4.5 para o *Weka* (HALL et al., 2009).

nenhuma das classes. Esses dados foram compostos por frequências e métodos estatísticos como: IM, χ^2, Teste-T e coeficiente Dice. O treinamento foi executado com 1136 instâncias, sendo metade com EMs verdadeiras (tomando como base as 568 EMs do *Gold Standard*) e a outra metade com EMs falsas. Após combinar de várias maneiras os métodos estatísticos, o melhor resultado encontrado para a classificação ocorreu com a utilização dos métodos IM, χ^2 e Dice. Portanto, este modelo foi utilizado para os dados de teste, que continham 15.000 candidatas à EM. O resultado apresentado pelo J48 foi de 12.782 bigramas classificados como EM e este consiste no quinto dicionário, chamado de **D5**.

4.3.4 Manual

Também foram criados dois dicionários de EMs de forma manual. Dicionários manuais geralmente exigem um custo alto para criação, por necessitarem de um especialista e também pelo tempo elevado em sua construção. O objetivo da geração deste tipo de dicionário para este trabalho, é avaliar a diferença performance de um dicionário construído manualmente, em relação a outro automaticamente gerado, sujeito a ruídos. As EMs deste dicionário foram selecionadas a partir do conteúdo dos 310 tópicos de consulta. Para primeiro dicionário, a escolha das EMs foi baseada na seleção de bigramas que poderiam obter um significado diferenciado, caso as palavras estivessem unidas. Esse dicionário, o sexto criado, foi chamado de **D6** e é composto de 254 expressões. O segundo dicionário manual foi criado a partir do conhecimento de um especialista. Foi pedido para um linguista que classificasse, de acordo com seu conhecimento, como verdadeira ou falsa uma lista de possíveis candidatas à EMs, extraída a partir dos tópicos de consulta. Suas escolhas verdadeiras formaram o sétimo dicionário, chamado de **D7** e composto por 178 bigramas.

4.4 Criação de Índices

Para a execução do experimento, modificou-se o corpus de diferentes maneiras, através da utilização dos dicionários previamente criados (Seção 4.3). As EMs presentes em cada dicionário foram inseridas no corpus na forma de um termo único, formado pela junção das palavras que compõem a expressão. Para cada dicionário, foram adicionadas todas as EMs nele contidas e em locais apropriados, ou seja, ao final de documentos onde ocorressem a EM. Este corpus modificado é então indexado no sistema de RI e os índices criados estão descritos conforme a Tabela 4.2. Na Tabela 4.3 pode-se observar a quantidade de bigramas que compõem cada dicionário de EMs.

Tabela 4.2: Índices

Nome do Índice	**Sigla**	**Dicionário Inserido**
Baseline	BL	-
Compostos Nominais	CN	D1
Melhores Compostos Nominais	MCN	D2
Piores Compostos Nominais	PCN	D3
Gold Standard	GS	D4
Árvore de Decisão	AD	D5
Manual 1	M1	D6
Manual 2	M2	D7

Fonte: Próprio autor.

Tabela 4.3: Total EMs por Dicionário

Dicionário	**Bigramas**
D1	15.001
D2	7.500
D3	17.328
D4	529
D5	12.782
D6	254
D7	178

Fonte: Próprio autor.

Como mencionado, para fins de indexação, cada EM inserida ao final de cada texto, seja ele um documento ou um tópico, é unida com um traço inferior (*underscore*). Isso faz com que essas palavras se tornem um termo único no sistema de RI. A Figura 4.2 exemplifica como ficam representadas as EMs identificadas nos documentos ou tópicos.

Figura 4.2: Documentos com Inserção de EMs

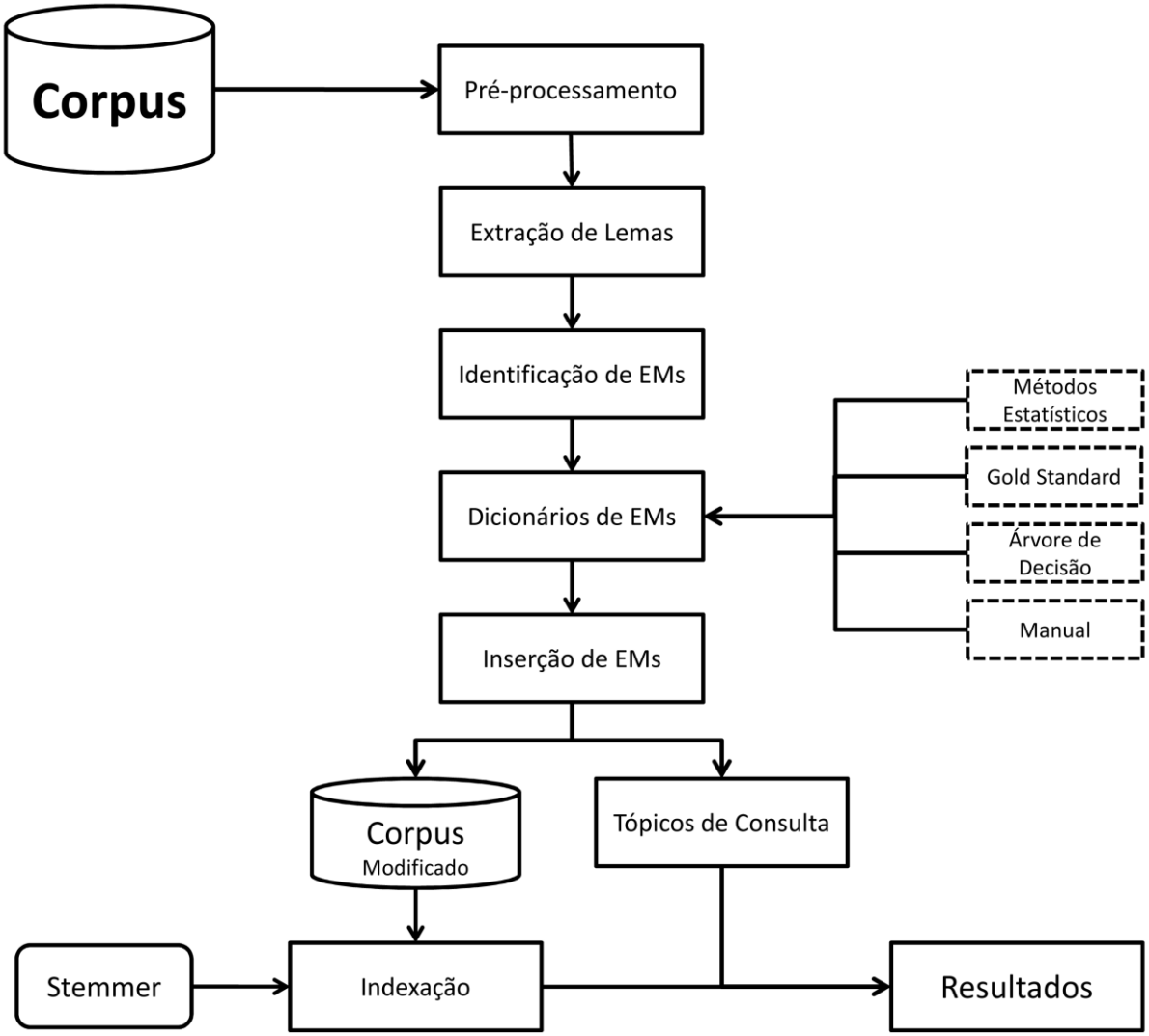
```
Tópico Original
- What was the role of the Hubble telescope in proving the
  existence of black holes?

Tópico Modificado
- what be the role of the hubble telescope in prove the
  existence of black hole ? black_hole
```

Fonte: Próprio autor.

4.5 Arquitetura

De maneira geral é possível definir, graficamente e de maneira genérica, o fluxo de dados e dos recursos utilizados por este trabalho, através da Figura 4.3. Inicialmente, um pré-processamento é executado no corpus original e são extraídos apenas os lemas dos termos existentes. A partir dos lemas são identificados os bigramas, formados por sequências de compostos nominais, e estes formam o que é chamado de dicionários de EMs. Os dicionários são criados a partir de diferentes métodos, sejam estatísticos, de uma listagem de EMs pré-selecionada em um dicionário, classificados por uma árvore de decisão ou mesmo manualmente. Após, esta lista de termos selecionada é utilizada no corpus e nos tópicos de consulta, em documentos que contenham EMs idênticas as presentes do dicionário. Após a inserção, o corpus modificado é indexado e os resultados são obtidos através dos tópicos de consulta, de onde é gerado um *ranking* dos documentos considerados relevantes para determinada consulta. Por fim, a análise dos resultados é dada pelos resultados gerados pelas métricas de avaliação.

Figura 4.3: Arquitetura



Fonte: Próprio autor.

5 EXPERIMENTOS

Neste capítulo é descrito o experimento realizado com o intuito de avaliar o impacto da inserção de Expressões Multipalavras nos resultados gerados pelo sistema de Recuperação de Informação. As análises do experimento foram divididas em três etapas: (A) Conjunto Total de Tópicos (Seção 5.1), onde é dada uma visão geral dos efeitos da inserção de EMs; (B) Tópicos Modificados por Expressões Multipalavra (Seção 5.2), onde são avaliados apenas os tópicos de consulta que contém EMs. Por fim, na terceira avaliação (C) é feito um comparativo dos resultados entre os diferentes tipos de índices criados com o intuito de verificar a contribuição da qualidade dos dicionários (Seção 5.3).

5.1 Avaliação A – Conjunto Total de Tópicos

Esta primeira avaliação teve como objetivo, de uma forma geral, investigar os efeitos da inserção de EMs nos documentos e nos tópicos de consulta. Após a indexação, para cada tipo de índice gerado, foram também inseridas EMs nos tópicos de consulta, sempre de acordo com os dicionários utilizados para cada índice[3].

Com as oito variações do corpus foram obtidos resultados individuais para cada um desses índices. O resultado apresentado na Tabela 5.1 resume-se através do número de documentos relevantes retornados, pela precisão média e pelo *R-Precision*. Ao todo, existem 6.379 documentos relevantes para os 310 tópicos de consulta.

Tabela 5.1: Resultados – Avaliação A

Nome do Índice	Rel. Retornados	Precisão Média	*R-Precision*
Baseline	3.967	0.1170	0.1287
Compostos Nominais	4.007	0.1179	0.1294
Melhores Compostos Nominais	3.972	0.1156	0.1271
Piores Compostos Nominais	3.982	0.1150	0.1252
Gold Standard	3.980	0.1193	0.1314
Árvore de Decisão	4.002	0.1178	0.1299
Manual 1	4.064	0.1217	0.1386
Manual 2	4.044	0.1205	0.1354

Fonte: Próprio autor.

[3] Para o índice *Baseline*, utilizou-se os tópicos de consulta sem nenhuma modificação.

Analisando a tabela é possível observar uma pequena vantagem nos índices Manual 1 e Manual 2, nos três tipos de resultados gerados. Uma explicação plausível pode ser o fato das escolhas de candidatas à EMs terem sido feitas a partir do conteúdo dos tópicos e não dos documentos, como é feito nos índices que utilizam compostos nominais (CN, MCN, PCN) retirados dos documentos, por exemplo. A seguir, o melhor resultado é do índice *Gold Standard*, o que a princípio era esperado por serem consideradas EMs verdadeiras, o que acarretaria em melhores resultados. Logo após, o índice de Compostos Nominais, com uma sutil vantagem sobre o *Baseline*. O *Baseline*, de forma inesperada, obteve um resultado melhor que os Melhores e Piores Compostos Nominais, pois não estava prevista uma possível perda nos resultados com a inserção de EMs. Porém, ao se comparar o ganho ou a perda individual de cada tópico foi verificado que, por exemplo, o índice MCN comparado ao *Baseline*, de 310 tópicos obteve ganho em 149 e perda em 108 casos. Desses 149 casos, a média de ganho da precisão média é de 0,0124 e a média de perda é de 0,0145, ou seja, pela média de perda ser maior é entendido o motivo pelo qual a precisão média do *Baseline* obter um resultado melhor que o MCN e PCN.

Entretanto, como esses valores de ganho e perda da precisão média são muito baixos, optou-se então por calcular um percentual de 5% (5 pontos percentuais) entre os resultados a serem comparados, para que sejam considerados relevantes ou não (BUCKLEY; VOORHEES, 2000). Por exemplo, a diferença entre os valores resultantes de dois índices diferentes para um mesmo tópico deve ser 5% maior ou 5% menor para que seja considerado um ganho ou uma perda significante. Dessa forma, é possível ver nas Tabelas 5.2 e 5.3 que os índices MCN e PCN melhoram em relação ao *Baseline*, novamente em uma comparação individual entre os tópicos.

Tabela 5.2: *Baseline* x MCN

	N.º Tópicos	**%**
Ganho	60	19,35%
Perda	35	11,29%
Não Relevante	215	69,35%
Total	310	100,00%
Diferença entre ganho e perda		8,06%

Fonte: Próprio autor.

Tabela 5.3: *Baseline* x PCN

	N.º Tópicos	**%**
Ganho	26	8,39%
Perda	21	6,77%
Não Relevante	263	84,84%
Total	310	100,00%
Diferença entre ganho e perda		1,61%

Fonte: Próprio autor.

No caso do MCN, o ganho é de quase 20% dos casos e o PCN, tem a diferença entre o ganho e a perda menor que 2%, o que vai ao encontro da hipótese investigada por este trabalho, de que melhores EMs resultam em uma melhor recuperação.

Por fim, foi concluído nesta primeira análise, a necessidade de se fazer uma avaliação mais profunda dos tópicos que possuem Expressões Multipalavras. Isso porque existe a possibilidade de que a inserção de EMs seja prejudicial ao resultado de tópicos que não possuem EMs inseridas. Um tópico de consulta que não possua EM retorna um documento relevante na primeira posição de seu ranking e, ao inserir EMs nos documentos, pode acarretar que esse documento da primeira posição tenha uma EM adicionada ao seu conteúdo, reduzindo assim o peso de outros termos, prejudicando o resultado final. Isso ocorre, por exemplo, no tópico de consulta 344, onde o escore de um documento relevante em relação ao índice BL diminui ao inserir, neste documento, uma EM *(home team)* que foi adicionada durante a criação do índice MCN, de *0.470566* para *0.455250*.

5.2 Avaliação B – Tópicos Modificados por Expressões Multipalavra

Esta avaliação teve como objetivo o estudo aprofundado dos efeitos causados na recuperação de documentos de tópicos de consulta em que houve a inserção de Expressões Multipalavras. Para isso, foram utilizados os mesmos índices anteriores e foi feita uma avaliação individual dos tópicos de consulta, para que se pudesse obter um maior entendimento dos casos em que a adição de expressões melhorou ou piorou os resultados.

Cada dicionário de EMs foi criado de maneira diferente, além das expressões neles contidos variarem, a quantidade de expressões que formam cada dicionário também são diferentes.

Como mencionado anteriormente, no total dos 310 tópicos de consulta, para cada índice criado, foram inseridas EMs de apenas um dos dicionários. Isso modificou diferentes tópicos, pois cada dicionário contém uma lista diferente de possíveis EMs. A Tabela 5.4 demonstra a quantidade de tópicos alterados em cada índice, de acordo com o dicionário utilizado.

Primeiramente, é interessante observar os valores de precisão média de todos os tópicos que sofreram alterações pela adição de EMs. Esses valores são mostrados pela Tabela 5.5.

Tabela 5.4: Tópicos com Expressões Multipalavras

Nome do Índice	Total de EMs	Tópicos c/ EM	% Alterado
Baseline	0	0	0,00%
Compostos Nominais	15.001	75	24,19%
Melhores Compostos Nominais	7.500	41	13,23%
Piores Compostos Nominais	17.328	28	9,03%
Gold Standard	529	9	2,90%
Árvore de Decisão	12.782	51	16,45%
Manual 1	254	195	62,90%
Manual 2	178	152	49,03%
Total de Tópicos	-	310	100,00%

Fonte: Próprio autor.

Tabela 5.5: Resultados – Avaliação B

Índice	Precisão Média
Compostos Nominais	0.1011
Melhores Compostos Nominais	0.0939
Piores Compostos Nominais	0.1224
Gold Standard	0.2393
Árvore de Decisão	0.1193
Manual 1	0.1262
Manual 2	0.1236

Fonte: Próprio autor.

Conforme a Tabela 5.5 foi verificado que o índice do *Gold Standard* obteve um valor bem acima dos demais, de certa forma previsto por se tratarem de Expressões Multipalavras consideradas verdadeiras. Após, foram avaliados os índices com escolhas manuais de EMs, Manual 1 e Manual 2, respectivamente. Neste caso, o que se considera como os Piores Compostos Nominais, obteve resultados gerais superiores aos restantes como, Árvore de Decisão, Compostos Nominais e Melhores Compostos Nominais, nesta ordem. Um dos motivos para esse resultado, talvez se deva ao fato de quantidade de EMs inseridas ser maior do que os outros índices em questão, mesmo que tenha alterado um número menor de tópicos de consulta. É importante lembrar que, mesmo que altere poucos tópicos de consulta, essa maior quantidade de EMs altera também diversos documentos do corpus em questão.

Antes de detalhar os motivos que causaram esses resultados, podemos observar as Tabelas de 5.6 até 5.12. Elas mostram a quantidade de tópicos que melhoraram ou pioraram, comparando-se com o *Baseline*, e os que não tiveram diferenças relevantes, utilizando os mesmos 5% de tolerância da análise anterior.

Tabela 5.6: *Baseline* x CN

Ganho	34	45,33%
Perda	10	13,33%
Não Relevante	31	41,33%
Total	75	100,00%
Diferença entre ganho e perda		32,0%

Fonte: Próprio autor.

Tabela 5.7: *Baseline* x MCN

Ganho	19	46,34%
Perda	6	14,63%
Não Relevante	16	39,02%
Total	41	100,00%
Diferença entre ganho e perda		31,7%

Fonte: Próprio autor.

Tabela 5.8: *Baseline* x PCN

Ganho	9	32,14%
Perda	2	7,14%
Não Relevante	17	60,71%
Total	28	100,00%
Diferença entre ganho e perda		25,0%

Fonte: Próprio autor.

Tabela 5.9: *Baseline* x GS

Ganho	5	55,56%
Perda	2	22,22%
Não Relevante	2	22,22%
Total	9	100,00%
Diferença entre ganho e perda		33,3%

Fonte: Próprio autor.

Tabela 5.10: *Baseline* x AD

Ganho	22	43,14%
Perda	5	9,80%
Não Relevante	24	47,06%
Total	51	100,00%
Diferença entre ganho e perda		33,3%

Fonte: Próprio autor.

Tabela 5.11: *Baseline* x Manual 1

Ganho	80	41,03%
Perda	37	18,97%
Não Relevante	78	40,00%
Total	195	100,00%
Diferença entre ganho e perda		22,1%

Fonte: Próprio autor.

Tabela 5.12: *Baseline* x Manual 2

Ganho	61	40,13%
Perda	28	18,42%
Não Relevante	63	41,45%
Total	152	100,00%
Diferença entre ganho e perda		21,7%

Fonte: Próprio autor.

Nas tabelas anteriores, o Total indica a quantidade de tópicos alterados e o percentual corresponde ao ganho ou a perda. Conforme pode ser observado em todos os casos, a inserção de EMs melhorou mais do que piorou os resultados recuperados pelos tópicos, em termos quantitativos. O melhor caso de ganho foi com o índice *Gold Standard*, onde 55,56% dos tópicos melhoraram, porém esse mesmo índice obteve o maior percentual de perda, 22,22%. Analisando o PCN, é possível identificar que os Piores Compostos Nominais têm o menor ganho (relacionado ao *Baseline*) se comparado a todos os outros índices: 32,14%. Apesar de ter também a menor perda, 60,71% dos tópicos de consulta modificados pelo PCN não tiveram diferenças relevantes se comparados ao *Baseline*. Dessa forma, pode-se concluir que

o índice PCN é o que menos altera significativamente o resultado da recuperação. Os índices CN e MCN tiveram um resultado semelhante, sabendo que o dicionário utilizado para a criação do MCN é um subconjunto do dicionário utilizado pelo CN, conclui-se então que, pelos valores de ganho, a escolha de melhores candidatas à EM não prejudicam o percentual de acerto, apenas melhoram de forma sutil. Porém, é importante ressaltar que o custo computacional para a inserção dessas EMs no corpus foi reduzido pela metade. Em termos de percentual de ganho os índices M1 e M2 foram superiores apenas ao PCN, entretanto estão próximos dos demais resultados, incluindo o índice AD, que obteve um resultado intermediário aos manuais e ao CN.

Para ilustrar como acontecem os ganhos ou perdas, serão apresentados alguns tópicos de forma mais detalhada, começando pelo Tópico de Consulta 141 (Figura 5.1) e o índice CN. Este foi o melhor caso de ganho no resultado da recuperação entre todos os índices utilizados.

Figura 5.1: Tópico de Consulta 141

```
<num>141</num>
<title>
letter bomb for kiesbauer find information on the explosion of a
letter bomb in the studio of the tv channel pro7 presenter
arabella kiesbauer . letter_bomb letter_bomb tv_channel
</title>
```

Fonte: Próprio autor.

Em uma análise das 10 primeiras posições ranqueadas pelo escore, pode-se observar que na Tabela 5.13 que o documento relevante (em negrito) está na quarta posição do ranking do *Baseline*. Ao inserir as expressões "*letter bomb*" (em português, carta-bomba), duas vezes, pois ela ocorre duas vezes no tópico original, e *"tv channel"* (em português, canal de TV) que estavam no dicionário D1, utilizada pelo índice CN, este documento relevante aumenta seu escore e retorna na primeira posição. A precisão média neste tópico aumentou 75 pontos percentuais, de 0,2500 no *Baseline* para 1,000, no índice CN. Também é possível observar que o documento que estava na primeira posição no ranking indicado pelo *Baseline* diminuiu seu escore e ficou classificado na quarta posição do ranking indicado pelo CN. Este documento continha informações sobre uma "pequena bomba localizada no lado de fora da embaixada russa", não sendo relevante para o Tópico 141 e sendo corretamente relegado a uma posição mais baixa.

Tabela 5.13: Ranking Tópico 141 - *Baseline*

Posição	Documento	Escore
P1	LA043094-0230	0.470900
P2	GH950823-000105	0.459994
P3	GH951120-000182	0.439536
P4	**GH950610-000164**	**0.430784**
P5	GH950614-000122	0.428766
P6	LA091894-0425	0.428429
P7	GH950829-000082	0.422941
P8	GH950220-000162	0.411968
P9	GH950318-000131	0.406006
P10	GH950829-000037	0.402806

Fonte: Próprio autor.

Tabela 5.14: Ranking Tópico 141 - CN

Posição	Documento	Escore
P1	**GH950610-000164**	**0.457950**
P2	GH950614-000122	0.436753
P3	GH950823-000105	0.423938
P4	LA043094-0230	0.421757
P5	GH951120-000182	0.400123
P6	GH950829-000082	0.393195
P7	LA091894-0425	0.386613
P8	GH950705-000100	0.384116
P9	GH950220-000162	0.382157
P10	GH950318-000131	0.380471

Fonte: Próprio autor.

Um fato interessante sobre esse tópico é que apenas a EM *"letter bomb"*, influencia no resultado. Isso foi comprovado pois no índice MCN, cujo dicionário não possui essa EM em questão, o tópico foi alterado por conta da EM *"tv channel"* e não houve ganho, nem perda no resultado. Já nos índices *Gold Standard* e AD, apenas *letter bomb* foi adicionado e o resultado foi o mesmo ganho citado anteriormente. Já os índices M1 e M2, obtiveram o mesmo ganho, porém utilizando tanto *"letter bomb"*, quanto *"tv channel"*.

O segundo ganho mais alto foi do índice M1, no Tópico 173. A melhora foi de 0,2841 pontos, passando de 0,2159 no *Baseline* para 0,5000 no M1. A Tabela 5.15 mostra os cinco melhores resultados ocasionados pela inserção de EMs. Este ganho foi calculado pela diferença dos valores obtidos pelos índices com EMs e o *Baseline*. Na coluna *EM*, o valor entre parênteses indica a quantidade de vezes que a EM ocorre no tópico.

Tabela 5.15: Os Cinco Melhores Resultados

Posição	Índice	Ganho	Tópico	EM
1	CN, GS, AD, M1 e M2	+0,7500	141	*letter bomb(2) e tv channel*
2	M1	+0,2841	173	*top quark(2)*
3	M1 e M2	+0,1943	233	*global warming(2) +2 EMs*[4]
4	CN, MCN, e M2	+0,12xx[5]	190	*child labor(2)*
5	CN, MCN e M1	+0,09xx[4]	3180	*sex education(2)*

Fonte: Próprio autor.

Por outro lado, foi encontrado um ponto negativo dos índices M1 e M2, esses, apesar de terem melhorado mais do que piorado os resultados dos tópicos de consulta, atingiram

[4] *climate change*(2) e *greenhouse effect*(3).

[5] xx, pois os três índices tiveram um resultado final diferente devido a inserção de outras EMs que interferiram de certa forma o resultado.

valores muito altos de perda como: -0,2734, -0,1500 e -0,1489 em tópicos que continham as EMs *"human being"*, *"mobile phone"* e "legal action", respectivamente. Foi verificado que esses valores ocorreram apenas nos índices de dicionários manuais, que extraíram EMs selecionadas a partir dos bigramas existentes nos tópicos de consulta. Já os outros índices tiveram como sua maior perda os resultados apresentados na Tabela 5.16. Da mesma forma que a tabela de ganhos, o resultado foi calculado pela diferença de cada índice comparado ao *Baseline*.

Tabela 5.16: Piores Resultados

Índice	**Perda**	**Tópico**	EM
CN	-0,0680	184	*maternity leave(2)*
MCN	-0,0123	158	*soccer match*
PCN	-0,0415	74	*tunnel project*
GS	-0,0399	230	*space shuttle*
AD	-0,0337	230	*space shuttle*

Fonte: Próprio autor.

Como é possível observar, excluindo-se os índices manuais a maior perda encontrada é de -0,0680, no Tópico 184 (Figura 5.2) do índice CN. Já o MCN obteve o menor valor de perda entre os índices, -0,0123 no Tópico 158 que contém a expressão "soccer match".

Figura 5.2: Tópico de Consulta 184

```
<num>184</num>
<title>
maternity leave in europe find document that give information on
provision concern the lenght of maternity leave in europe .
maternity_leave maternity_leave
</title>
```

Fonte: Próprio autor.

Nas Tabelas 5.17 e 5.18 é observado que o documento relevante (em negrito) que estava na primeira colocação no *Baseline*, cai para a quarta colocação com a inserção de EMs e com o escore mais baixo. Porém, é importante ressaltar que o documento *GH950222-000131* contém a expressão "maternity leave", entretanto o documento não é considerado relevante para esse tópico. Por outro lado, apesar da expressão existir na consulta, o documento *GH951107-000139*, considerado relevante para esse tópico de consulta, não contém essa expressão.

Tabela 5.17: Ranking Tópico 184 - *Baseline*

Posição	Documento	Escore
P1	**GH951107-000139**	**0.329321**
P2	LA061794-0102	0.314456
P3	LA010294-0122	0.308068
P4	GH950612-000094	0.303283
P5	GH950119-000154	0.300739
P6	**GH951108-000133**	**0.300260**
P7	GH951012-000163	0.299661
P8	GH950501-000133	0.296750
P9	GH951005-000119	0.294093
P10	LA102794-0371	0.290570

Fonte: Próprio autor.

Tabela 5.18: Ranking Tópico 184 - CN

Posição	Documento	Escore
P1	GH950222-000131	0.313635
P2	LA061794-0102	0.298905
P3	LA010294-0122	0.296016
P4	**GH951107-000139**	**0.295480**
P5	GH950502-000121	0.294713
P6	**GH951108-000133**	**0.293321**
P7	GH950612-000094	0.291418
P8	GH950119-000154	0.288973
P9	GH951020-000109	0.286908
P10	GH950501-000133	0.285141

Fonte: Próprio autor.

Com essas análises, é possível concluir que, de fato, a inserção de EMs aumentam as chances de documentos que contenham a mesma expressão sejam retornados em posições melhores ranqueadas, diferentemente do caso não em que não fossem indexadas. Porém, nem sempre a existência de uma EM no documento garante que ele seja relevante para determinado tópico que também a contenha. Para resolver estes casos, seria necessário a utilização de uma análise semântica para estas consultas, podendo assim definir sua relevância para o tópico. Outro ponto importante, é que mesmo com o ganho, em alguns casos analisados, além do aumento do escore dos documentos relevantes, o escore dos não relevantes também é incrementado, evitando assim um ganho ainda maior dos resultados.

5.3 Avaliação C – Comparativo entre os índices

Esta última avaliação teve como objetivo a comparação direta entre os índices, diferentemente das avaliações anteriores, onde os índices eram comparados apenas com o *Baseline*. Como cada índice modifica tópicos de consulta que podem ser diferentes dos que são modificados por outros índices (por utilizarem dicionários diferentes), optou-se por fazer uma análise geral, ou seja, dos 310 tópicos de consulta utilizados no experimento. Da mesma maneira das avaliações anteriores, os comparativos levaram em consideração a diferença de 5% na precisão média entre os resultados de cada índice para cada tópico, considerando assim, o resultado relevante ou não. Dentre as várias comparações possíveis, foram selecionadas as mais relevantes para esta avaliação.

Primeiramente, ao se analisar os índices de Compostos Nominais (CN, MCN e PCN), podemos observar nas Tabelas 5.19 e 5.20 que o comparativo entre o índice CN é superior em

ambos os casos. Entretanto, a diferença de 5 tópicos (1,61 ponto percentual) (Tabela 5.19), é muito sutil em comparação com o MCN, que consiste nas "melhores" EMs do índice CN, de acordo com os métodos de extração e classificação de EMs utilizados pelo presente trabalho. Na comparação com o PCN (Tabela 5.20) essa diferença é maior, sendo 19 tópicos (6,13 pontos percentuais) a mais a favor do CN, confirmando, de certa forma, o fato da escolha de melhores candidatas à EM retornarem melhores resultados nas consultas.

Tabela 5.19: CN x MCN

Índice	**N.º Tópicos**	**%**
CN	39	12,58%
MCN	34	10,97%
Não Relevante	237	76,45%
Total	310	100,00%
Diferença entre CN e MCN		1,61%

Fonte: Próprio autor.

Tabela 5.20: CN x PCN

Índice	**N.º Tópicos**	**%**
CN	67	21,61%
PCN	48	15,48%
Não Relevante	195	62,90%
Total	310	100,00%
Diferença entre CN e PCN		6,13%

Fonte: Próprio autor.

Também pode ser observado na Tabela 5.21, uma pequena superioridade para o índice MCN em relação ao PCN, de aproximadamente 3 pontos percentuais (nove tópicos). Lembrando que a diferença entre EMs inseridas entre esses índices é de pouco menos de 10.000 expressões, podendo concluir assim que a qualidade das expressões dos dicionários é mais relevante do que a quantidade.

Tabela 5.21: MCN x PCN

Índice	**N.º Tópicos**	**%**
MCN	46	14,84%
PCN	37	11,94%
Não Relevante	227	73,23%
Total	310	100,00%
Diferença entre MCN e PCN		2,90%

Fonte: Próprio autor.

Na comparação entre o MCN e o AD (Tabela 5.22), a diferença apresentada entre eles é praticamente insignificante, de menos de 1 ponto percentual (ou dois tópicos). Porém, como mencionado também nas avaliações comparativas anteriores do MCN, a quantidade de EMs inseridas no índice AD é amplamente maior.

Tabela 5.22: MCN x AD

Índice	N.º Tópicos	%
MCN	40	12,90%
AD	38	12,26%
Não Relevante	232	74,84%
Total	310	100,00%
Diferença entre MCN e AD		0,65%

Fonte: Próprio autor.

Entretanto, o índice MCN é inferior aos resultados dos índices M1 e M2 (Tabelas 5.23 e 5.24), apesar desses índices serem gerados com uma quantidade menor de EMs: 5,81 pontos percentuais ou 18 tópicos, comparado ao M1 e 2,26 pontos percentuais ou sete tópicos, em relação ao M2. É relevante afirmar que, caso fossem omitidos os critérios de 5% acima e abaixo para que se considere um resultado relevante, a maioria dos tópicos cuja precisão média é mais alta seria do índice MCN. Portanto, podemos concluir que na quantidade o MCN levaria vantagem, porém sua qualidade no ganho dos tópicos é inferior aos índices M1 e M2.

Tabela 5.23: MCN x M1

Índice	N.º Tópicos	%
MCN	66	21,29%
M1	84	27,10%
Não Relevante	160	51,61%
Total	310	100,00%
Diferença entre MCN e M1		5,81%

Fonte: Próprio autor.

Tabela 5.24: MCN x M2

Índice	N.º Tópicos	%
MCN	66	21,29%
M2	73	23,55%
Não Relevante	171	55,16%
Total	310	100,00%
Diferença entre MCN e M2		2,26%

Fonte: Próprio autor.

E por fim, a Tabela 5.25 indica o índice MCN, comparado ao índice *Gold Standard*. É possível observar que o índice GS é inferior neste caso (13 tópicos ou 4,19 pontos percentuais), diferente do que pode ser observado na Avaliação B (Seção 5.2), onde o índice GS foi o que obteve o melhor resultado. Isso nos permite entender que os valores retornados dos tópicos de consulta para o índice GS quando melhoram, melhoram bastante e quando pioram, pioram pouco. Porém, em uma comparação com o índice MCN na quantidade de tópicos que melhoram, seu resultado acaba sendo inferior, também atribuído pelo fato do índice GS ter 1/3 da quantidade de EMs do índice MCN.

Tabela 5.25: GS x MCN

Índice	**N.º Tópicos**	**%**
GS	46	14,84%
MCN	59	19,03%
Não Relevante	205	66,13%
Total	310	100,00%
Diferença entre GS e MCN		4,19%

Fonte: Próprio autor.

Em resumo, conclui-se através dessa avaliação que a escolha de melhores candidatas a Expressão Multipalavra levam, como esperado, à melhores resultados. A seleção de melhores candidatas à EMs, obtidas através dos métodos de extração utilizados por este trabalho, melhoraram o resultado da recuperação de documentos relevantes relacionados aos tópicos de consultas. Entretanto, quanto maior o número de EMs utilizadas, mais sutil é o ganho. Comprovou-se então que, neste caso, a qualidade das EMs utilizadas é superior à quantidade EMs, isso ocorreu principalmente na comparação dos Melhores Compostos Nominais com os índices Manuais. Já na comparação dos índices MCN com o GS, fica evidente que a extração de candidatas à EMs do próprio corpus, colaboram para uma pequena vantagem em relação à dicionários de EMs previamente elaborados.

6 CONCLUSÕES E TRABALHOS FUTUROS

O trabalho apresentado consiste em explorar a área das Expressões Multipalavras, desde seu conceito até suas diferentes classes, visando investigar eventuais melhorias da indexação destas expressões em sistemas de Recuperação de Informação. Para isso, foi necessária uma avaliação de quais técnicas de extração de EMs, melhor se adaptam a determinados tipos de expressões e corpora (ACOSTA; VILLAVICENCIO; MOREIRA, 2011). Em um sistema real, a indexação de termos atômicos pode causar uma perda semântica e para que a recuperação possa ser mais precisa e eficiente, os termos de indexação devem corresponder aos conceitos presentes nos documentos. Sendo assim, torna-se necessário identificar tais termos, a fim de que o sistema de RI os trate de maneira adequada durante a recuperação.

Com o desenvolvimento deste trabalho foi possível avaliar a contribuição de Expressões Multipalavras em sistemas de Recuperação de Informação. De uma forma geral, pode-se afirmar que a inserção de EMs tende a melhorar os resultados alcançados por mecanismos de busca de um SRI.

Observando os resultados obtidos neste trabalho, percebe-se que na grande maioria dos casos, ao se adicionar EMs tanto nos tópicos, quanto nos documentos, a recuperação de documentos relevantes aumentou consideravelmente. Isso se deve à escolha de EMs verdadeiras e que ocorrem também com certa frequência no corpus. Também é possível concluir que a qualidade das EMs utilizadas influencia mais o resultado do que a quantidade. A melhora nos resultados, a partir de certo número de EMs verdadeiras, começa a se tornar sutil. Isso ficou comprovado com os resultados similares do índice dos Melhores Compostos Nominais (MCN) em relação ao índice dos Compostos Nominais (CN) na Avaliação B, da mesma forma que o índice dos Piores Compostos Nominais (PCN) teve o menor ganho nessa mesma avaliação. Outro indício encontrado é o fato do índice *Gold Standard* (GS) estar sempre entre os melhores resultados nas avaliações, mesmo tendo um número consideravelmente pequeno de EMs.

Entretanto, em alguns casos é possível associar a perda nos resultados à maneira em como se avalia a relevância de um documento. Um documento pode ser considerado relevante se possui informações qualificadas sobre a necessidade do usuário em determinado momento. Como exemplo, podemos citar um dos tópicos de consulta utilizado por este trabalho. Nele são requisitadas informações sobre a causa da morte do vocalista de uma banda de rock e documentos considerados relevantes são aqueles que não só informem sua morte, mas

também citem a causa do óbito. Em outros casos, alguns tópicos de consulta exigem um processamento mais aprofundado, como uma análise semântica para que se determine a relevância do documento (por exemplo, a recuperação de documentos que sejam favoráveis ao presidente dos Estados Unidos da América). Porém, esse é um problema geral de sistemas de RI e não somente deste trabalho.

Por fim, comprovou-se através das avaliações do experimento a hipótese da identificação e tratamento adequado das Expressões Multipalavras para posteriormente indexá-las como uma unidade única em um mecanismo de indexação, melhoram de certa forma os resultados de um Sistema de Recuperação de Informação.

Como trabalhos futuros, propõe-se utilizar outros tipos de EMs e não apenas Compostos Nominais, como foram utilizados neste trabalho. Novos métodos de extração e um estudo mais aprofundado em Entidades Nomeadas são bons temas complementares ao assunto aqui abordado. Outro ponto interessante para outra avaliação seria a variação de corpora, diferente dos textos jornalísticos, pois cada domínio possui terminologias peculiares. Além disso, testes com outros modelos de RI também podem ser efetuados visando estudar se estes podem ocasionar um comportamento diferente do vetorial, utilizado por este trabalho.

REFERÊNCIAS

ACOSTA, O.; GERALDO, A.; ORENGO, V. M.; VILLAVICENCIO, A. UFRGS@CLEF2008: Indexing Multiword Expressions for Information Retrieval**.** In: **Working Notes of the Workshop of the Cross-Language Evaluation Forum - CLEF**. Aarhus, Denmark: [s.n.], 2008.

ACOSTA, O. C.; VILLAVICENCIO, A.; MOREIRA, V. P. Identification and treatment of multiword expressions applied to information retrieval. In: **Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (MWE '11), p. 101–109. ISBN 978-1-932432-97-8.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. [S.l.]: ACM Press / Addison-Wesley, 1999.

BALDWIN, T.; BANNARD, C.; TANAKA, T.; WIDDOWS, D. An empirical model of multiword expression decomposability. In: **Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (MWE '03), p. 89–96.

BALDWIN, T.; VILLAVICENCIO, A. Extracting the unextractable: A case study on verb-particles. In: **Proceedings of the 6th Conference on Natural Language Learning - Volume 20**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (COLING-02), p. 1–7.

BERRY-ROGGHE, G. L.; WULZ, H. An overview of plidis, a problem solving information system with german as query language. In: **Natural Language Communication with Computers.** London, UK, UK: Springer-Verlag, 1978. p. 87–132. ISBN 3-540-08911-X.

BLAHETA, D.; JOHNSON, M. Unsupervised learning of multi-word verbs. In: **Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations**. [S.l.: s.n.], 2001. p. 54–60.

BUCKLEY, C.; VOORHEES, E. M. Evaluating evaluation measure stability. In: **SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.** New York, NY, USA: ACM, 2000. p. 33–40. ISBN 1-58113-226-3.

CALZOLARI, N. et al. Towards best practice for multiword expressions in computational lexicons. In: **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002).** Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), 2002. ACL Anthology Identifier: L02-1259.

CAMBRIDGE International Dictionary of English. [S.l.]: Cambridge University Press, Cambridge ; New York, 1995. 1773 p. ISBN 0521484219 0521484693 0521482364.

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, 1990. ISSN 0891-2017.

CLEF. **Cross Language Evaluation Forum.** [S.l.]: TrebleCLEF Coordination Action, 2008. (Disponível em: < http://www.clef-campaign.org>. Acesso em: jul. 2011.

EVANS, D. A.; ZHAI, C. Noun-phrase analysis in unrestricted text for information retrieval. In: **Proceedings of the 34th Annual Meeting on Association for Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996. (ACL '96), p. 17–24.

EVERT, S.; KRENN, B. Using small random samples for the manual evaluation of statistical association measures. **Computer Speech and Language**, Academic Press Ltd., London, UK, UK, v. 19, n. 4, p. 450–466, oct. 2005. ISSN 0885-2308.

FELLBAUM, C. **WordNet: An Electronic Lexical Database.** Cambridge, MA: MIT Press, 1998. ISBN 978-0-262-06197-1.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA Data Mining Software: An Update**. [S.l.:s.n], 2009.

JACKENDOFF, R. **The Architecture of the Language Faculty**. [S.l.]: MIT Press, 1997. ISBN 0-262-60025-0.

JONES, K. S. What is the role of nlp in text retrieval? In: **Natural language information retrieval.** [S.l.]: University of Cambridge, 1997.

KITA, K.; KATO, Y.; OMOTO, T.; YANO, Y. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. **Journal of Natural Language Processing**, v. 1, n. 1, p. 21–33, 1994.

MANNING, C.; SHüTZE, H. **Foundations of Statistical Natural Language Processing**. [S.l.]: The MIT Press, 1999.

MILLER, G. A. Wordnet: a lexical database for english. **Commun. ACM**, ACM, New York, NY, USA, v. 38, p. 39–41, November 1995. ISSN 0001-0782.

NICHOLSON, J.; BALDWIN, T. Interpretation of compound nominalisations using corpus and web statistics. In: **Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties**. Sydney, Australia: Association for Computational Linguistics, 2006. p. 54–61.

ORENGO, V. M. **Assessing Relevance Using Automatically Translated Documents for Cross-Language Information Retrieval.** Thesis (PhD) — Middlesex University - London - UK, 2004.

PEARCE, D. Synonymy in collocation extraction. In: **Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations**. [S.l.:s.n], 2001.

PEARCE, D. A comparative evaluation of collocation extraction techniques. In: **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)**. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), 2002. ACL Anthology Identifier: L02-1169.

PORTER, M. F. **Readings in information retrieval**. In: JONES, K. S.; WILLETT, P. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. chp. An Algorithm for Suffix Stripping, p. 313–316. ISBN 1-55860-454-5.

QUINLAN, J. R. Induction of decision trees. **Mach. Learn**., Kluwer Academic Publishers, Hingham, MA, USA, v. 1, p. 81–106, March 1986. ISSN 0885-6125.

RAMISCH, C.; VILLAVICENCIO, A.; BOITET, C. Multiword expressions in the wild? the mwetoolkit comes in handy. In: **Coling 2010: Demonstrations.** Beijing, China: Coling 2010 Organizing Committee, 2010. p. 57–60.

ROBERTSON, S. E.; JONES, K. S. Relevance weighting of search terms. **Journal of the American Society for Information Science**, Wiley Subscription Services, Inc., A Wiley Company, v. 27, n. 3, p. 129–146, 1976. ISSN 1097-4571.

SAG, I.; BALDWIN, T.; BOND, F.; COPESTAKE, A.; FLICKIGER, D. A. et al. Multiword expressions: A pain in the neck for nlp. In: **Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing**. London, UK, UK: Springer-Verlag, 2002. (CICLing '02), p. 1–15. ISBN 3-540-43219-1.

SALTON, G. **The SMART Retrieval System - Experiments in Automatic Document Processing**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1971.

SALTON, G.; BUCKLEY, C. **Term Weighting Approaches in Automatic Text Retrieval**. Ithaca, NY, USA, 1987.

SALTON, G.; FOX, E. A.; WU, H. Extended boolean information retrieval. **Commun. ACM**, ACM, New York, NY, USA, v. 26, p. 1022–1036, November 1983. ISSN 0001-0782.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. [S.l.]: McGraw-Hill, 1983.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: **Proceedings of the International Conference on New Methods in Language Processing**. [S.l.: s.n.], 1994.

SHIMOHATA, S.; SUGIO, T.; NAGATA, J. Retrieving collocations by co-occurrences and word order constraints. In: **Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)**. Madrid, Spain: [s.n.], 1997. p. 476–81.

SMADJA, F. Retrieving collocations from text: Xtract. In: . [S.l.]: Computational Linguistics, 1993.

VILLAVICENCIO, A.; BOND, F.; KORHONEN, A.; MCCARTHY, D. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. **Computer Speech and Language**, Academic Press Ltd., London, UK, UK, v. 19, n. 4, p. 365–377, oct. 2005. ISSN 0885-2308.

VILLAVICENCIO, A.; RAMISCH, C.; MACHADO, A.; MEDEIROS CASELI, H. de;FINATTO, M. J. Identificação de expressões multipalavra em domínios específicos. **Linguamática**, v. 2, n. 1, p. 15–33, Abril 2010. ISSN 1647-0818.

WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.** San Francisco: Morgan Kaufmann, 2000.

WIVES, L. K.; OLIVEIRA, J. P. M. d. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. [S.l.], 2001.

ZETTAIR. The Zettair Search Engine. 2008.

ZHANG, Y. et al. Automated multiword expression prediction for grammar engineering. In: **Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (MWE '06), p. 36–44. ISBN 1-932432-84-1.

APÊNDICE <ÍNDICES X BASELINE SEM 5%>

Tabelas comparativas entre os índices e o *Baseline* em tópicos com EMs, sem a utilização dos 5% de relevância.

Tabela 1: *Baseline* x CN

Índice	Tópicos	%
Baseline	17	22,67%
CN	44	58,67%
Empate	14	18,67%
Total	75	100,00%

Fonte: Próprio autor.

Tabela 2: *Baseline* x MCN

Índice	Tópicos	%
Baseline	13	31,71%
MCN	21	51,22%
Empate	7	17,07%
Total	41	100,00%

Fonte: Próprio autor.

Tabela 3: *Baseline* x PCN

Índice	Tópicos	%
Baseline	9	32,14%
PCN	16	57,14%
Empate	3	10,71%
Total	28	100,00%

Fonte: Próprio autor.

Tabela 4: *Baseline* x GS

Índice	Tópicos	%
Baseline	2	22,22%
GS	6	66,67%
Empate	1	11,11%
Total	9	100,00%

Fonte: Próprio autor.

Tabela 5: *Baseline* x AD

Índice	Tópicos	%
Baseline	11	21,57%
AD	32	62,75%
Empate	8	15,69%
Total	51	100,00%

Fonte: Próprio autor.

Tabela 6: *Baseline* x Manual 1

Índice	Tópicos	%
Baseline	54	27,69%
M1	107	54,87%
Empate	34	17,44%
Total	195	100,00%

Fonte: Próprio autor.

Tabela 7: *Baseline* x Manual 2

Índice	Tópicos	%
Baseline	45	29,61%
M2	77	50,66%
Empate	30	19,74%
Total	152	100,00%

Fonte: Próprio autor.