UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

ANDRÉ MARANHÃO MACHADO

# Extração de Expressões Multipalavra em Corpora Técnicos

Projeto de Diplomação

Prof$^{a}$. Dr$^{a}$. Aline Villavicencio
Orientador

Porto Alegre, dezembro de 2009

# AGRADECIMENTOS

Agradeço à Aline pela orientação deste trabalho, à Rosa Viccari, Elder Santos e Tiago Primo que sempre me auxiliaram nos projetos de pesquisa em que atuei durante o curso. Agradeço também a minha família pelo apoio e aos amigos que conheci na faculdade pelos bons momentos compartilhados ao longo destes anos.

Este trabalho resultou de uma colaboração entre a Universidade Federal de São Paulo - São Carlos, a Universidade Federal do Rio Grande do Sul e a Universidade de Grenoble e foi descrito nos seguintes artigos:

- *Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains* publicado no *Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. Singapura, 2009.*
- *Identification of Multiword Expressions in Technical Domains: Investigating Statistical and Alignment-based Approaches* publicado no *7th Brazilian Symposium in Information and Human Language Technology.*
- *A Hybrid Approach for Multiword Expression Identification* aceito para a *International Conference on Computational Processing of Portuguese Language. 2010, Porto Alegre.*

# SUMÁRIO

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| BNC | British National Corpus |
| CVP | Construção verbo-partícula |
| EMP | Expressão multipalavra |
| MA | Medida de associação |
| MI | Mutual Information |
| PLN | Processamento de linguagem natural |
| PMI | Pointwise Mutual Information |
| POS | Part-of-Speech |
| PS | Poisson-Stirling |

# LISTA DE TABELAS

# RESUMO

Expressões multipalavra (EMPs) são um dos obstáculos para a obtenção de sistemas de PLN mais precisos. Particularmente, a falta de cobertura de EMPs em recursos lexicais pode impactar negativamente na performance de tarefas e aplicações, levando a perda de informação ou erros de comunicação. Isso é especialmente problemático em domínios técnicos, onde uma parte significativa do vocabulário é composta de EMPs. Este trabalho tem por objetivo investigar o uso de diferentes métodos para a identificação de EMPs em corpora técnicos. São usadas diversas fontes de dados, incluindo um *corpus* paralelo, utilizando textos em português e inglês de um *corpus* de Pediatria. Examina-se como uma segunda língua pode fornecer informações relevantes para essas tarefa. Este trabalho é uma extensão dos artigos abaixo:

- *Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains* publicado no *Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. Singapura, 2009.*
- *Identification of Multiword Expressions in Technical Domains: Investigating Statistical and Alignment-based Approaches* publicado no *7th Brazilian Symposium in Information and Human Language Technology.*
- *A Hybrid Approach for Multiword Expression Identification* aceito para o *International Conference on Computational Processing of Portuguese Language. 2010, Porto Alegre.*

**Palavras-chave:** Processamento de linguagem natural, expressões multipalavra, corpora paralelos, UFRGS.

**Extraction of Multiword Expressions in Technical Domains**

# ABSTRACT

Multiword Expressions (MWEs) are one of the stumbling blocks for more precise Natural Language Processing (NLP) systems. Particularly, the lack of coverage of MWEs in resources can impact negatively on the performance of tasks and applications, and can lead to loss of information or communication errors. This is especially problematic in technical domains, where a significant portion of the vocabulary is composed of MWEs. This work investigates the use of different approaches to the identification of MWEs in technical corpora. We look at the use of several sources of data, including a parallel corpus, using English and Portuguese data from a corpus of Pediatrics, and examining how a second language can provide relevant cues for this task. This is an extended version of the following papers:

- *Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains* published at the *Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. 2009, Singapore.*
- *Identification of Multiword Expressions in Technical Domains: Investigating Statistical and Alignment-based Approaches* published at the *7th Brazilian Symposium in Information and Human Language Technology. 2009, São Carlos.*
- *A Hybrid Approach for Multiword Expression Identification* accepted for the *International Conference on Computational Processing of Portuguese Language. 2010, Porto Alegre.*

**Keywords:** natural language processing, multiword expressions, parallel corpora, UFRGS.

# 1 INTRODUÇÃO

## 1.1 Motivação

O Processamento de Linguagem Natural (PLN) se trata de uma subárea multidisciplinar da Computação, que combina aspectos de Inteligência Artificial e Linguística Computacional, entre outras, para estudar os problemas de entendimento e geração de linguagens naturais. Uma aplicação de PLN é caracterizada pelo uso do algum conhecimento da linguagem natural (*e.g.* para que um computador seja capaz de gerar um sinal de áudio a partir de uma sequência de palavras (i.e., voz), é preciso que ele possua conhecimentos sobre fonética e fonologia). Desta forma, o PLN pode ser usado em áreas como verificação ortográfica e gramatical, sistemas de auxílio a leitura e escrita, extração e recuperação de informações, tradução automática, criação de resumos a partir de textos, sistemas de respostas a perguntas, entre outros.

O desempenho de aplicações de PLN depende de recursos lexicais eletrônicos — como dicionários, tesauros, ontologias — precisos e de alta cobertura. A geração desses recursos de forma manual é muito custosa e sujeita a erros em virtude da magnitude das linguagens naturais e de sua natureza dinâmica. Recursos lexicais para expressões multipalavra, em particular, são ainda escassos e de baixa cobertura. Assim, há a necessidade de se desenvolver métodos semiautomáticos para a geração desses recursos. Para tanto, são utilizadas grandes bases de textos, os corpora. Neste trabalho, são investigados métodos para identificação de EMPs a partir de um *corpus* de textos paralelos, isto é, um conjunto de textos equivalentes em diferentes idiomas.

Investiga-se, neste trabalho, técnicas de identificação de construções da linguagem conhecidas como *expressões multipalavra* (EMP) (do inglês *multiword expression*). EMPs podem ser definidas como qualquer combinação de palavras em que as propriedades sintáticas ou semânticas da expressão não podem ser obtidas de suas partes (SAG et al., 2002). De forma a mostrar a importância da identificação e tratamento adequado de EMPs, dois exemplos de aplicações de PLN são mostrados abaixo.

O primeiro exemplo é o `START Natural Language system`[1]. Esse sistema foi desenvolvido com o objetivo de responder a questões formuladas em linguagem natural (inglês). O sistema analisa as perguntas efetuadas e então busca uma resposta adequada em sua base de conhecimento. Atualmente, o sistema é capaz de responder a milhões de perguntas em linguagem natural a respeito de lugares, filmes, pessoas, etc. Desta forma, ao se realizar o questionamento *When did Elvis Presley die?* (Quando Elvis Presley morreu?), o sistema, após utilizar técnicas de PLN e re-

[1]Disponível em: `http://start.csail.mit.edu`

cuperação de informações, é capaz de responder adequadamente: *Elvis Presley died in 1977* (Elvis Presley morreu em 1977). Porém, ao se realizar a mesma pergunta substituindo-se o verbo *die* por uma expressão multipalavra equivalente como *kick the bucket* (bater as botas), o comportamento esperado do sistema não é observado (*Unfortunately, I don't know when Elvis Presley kicked the bucket.* — Infelizmente, eu não sei quando Elvis Presley chutou o balde, onde "chutou o balde" é interpretado de forma literal).

A falta de tratamento de EMPs pode, também, gerar problemas para sistemas de tradução automática, dois exemplos de traduções incorretas obtidas no `Yahoo Babelfish`[2] são mostrados na tabela 1.1.

Tabela 1.1: Resultados obtidos em tradutores automáticos quando do uso de EMPs

| **Expressão pesquisada** | **Tradução válida** | **Tradução obtida** |
|---|---|---|
| Who did spill the beans? | Quem abriu o bico? | Quem derramou os feijões? |
| The old horse has kicked the bucket. | O cavalo velho bateu as botas. | O cavalo velho retrocedeu a cubeta. |

Considerando a natureza flexível e dinâmica das EMPs e que o número dessas expressões na linguagem cotidiana é da mesma ordem de magnitude que o número de palavras únicas (JACKENDOFF, 1997), a identificação e tratamento adequado de EMPs se tornam críticos para um bom desempenho de aplicações de PLN.

## 1.2 Estrutura do trabalho

No presente trabalho, são investigadas e desenvolvidas técnicas de extração semiautomática de expressões multipalavra, baseadas na combinação de métodos estatísticos, de informações linguísticas e de alinhamento lexical, a partir de textos paralelos. Em particular, o foco está em identificar a ocorrência dessas expressões em textos paralelos de domínio especifico. Neste caso, são utilizados pares de texto português-inglês do domínio de pediatria, visto que a terminologia técnica consiste, em grande parte, de expressões multipalavra. Os métodos são avaliados de acordo com a performance na identificação de EMPs gerais e específicas do domínio de pediatria.

Neste contexto, este documento é estruturado da seguinte forma. No capítulo 2, é realizada uma revisão bibliográfica sobre o EMPs, onde são citados os trabalhos relacionados e as dificuldades encontradas para o seu tratamento. Além disso, são descritos métodos utilizados para a identificação de EMPs: métodos estatísticos, métodos baseados em alinhamento de textos paralelos e uma abordagem mista, que utiliza características dos métodos citados anteriormente. No capítulo 3, os recursos lexicais utilizados no trabalho são detalhados. Características do *corpus* paralelo são mostradas. É também descrito o processo de criação das listas de referência utilizadas no processo de avaliação. Após, no capítulo 4, os experimentos baseados nas abordagens descritas no capítulo 2 são descritos. Os resultados obtidos nos diferentes experimentos são discutidos e comparados. Por fim, no capítulo 5, são realizadas as considerações finais e os trabalhos futuros são propostos.

[2]Disponível em `http://babelfish.yahoo.com`

# 2 IDENTIFICAÇÃO DE EXPRESSÕES MULTIPALAVRA

## 2.1 Definição

Uma EMP pode ser definida como qualquer combinação de palavras em que as propriedades sintáticas ou semânticas da expressão não podem ser obtidas de suas partes (SAG et al., 2002). Exemplos de EMPs são verbos frasais da língua inglesa (*break down* — quebrar, romper), compostos nominais (*police car* — carro da polícia), *coffee machine* — máquina de café), expressões idiomáticas (*to rock the boat* — causar problemas), (*to let the cat out of bag* — abrir o bico, revelar um segredo), entre outros.

Essas expressões são muito numerosas, para BIBER et al. (1999), elas totalizam de 30% a 45% da língua inglesa falada e 21% de texto acadêmico e, para JACKENDOFF (1997), o número de EMPs no vocabulário de um falante nativo é da mesma ordem de magnitude que o número de palavras simples. Entretanto, estes valores estão possivelmente subestimados se considerarmos o vocabulário especializado utilizado em textos de domínio específico. Nesse tipo de texto, a terminologia utilizada consiste, em grande parte, de EMPs (*global warming* — aquecimento global, *protein sequencing* — sequenciamento de proteínas). Além disso, novas EMPs surgem constantemente (*weapons of mass destruction* — armas de destruição em massa, *axis of evil* — eixo do mal). Essas expressões são muito frequentes na linguagem coloquial e isso se reflete em diversas gramáticas e recursos lexicais existentes, onde quase metade das entradas são EMPs.

EMPs tem um papel importante em aplicações de PLN. Essas aplicações não devem apenas identificá-las mas também devem ser capazes de lidar com elas quando são encontradas (FAZLY; STEVENSON, 2007). Não as identificar nem as tratar adequadamente pode causar sérios problemas para muitas tarefas de PLN, especialmente aquelas envolvendo algum tipo de processamento semântico. Em uma avaliação do processo de análise sintática, por exemplo, BALDWIN et al. (2004) observou que, para uma amostra randômica de 20.000 sentenças do British National Corpus (BNC), mesmo com uma gramática de ampla abrangência para a língua inglesa, 8% do total de erros foram causados em decorrência de EMPs não existentes nos recursos lexicais utilizados.

Portanto, percebe-se uma grande necessidade de se obter formas semiautomáticas robustas capazes de adquirir informação lexical para EMPs de forma a aumentar significativamente a cobertura dos recursos utilizados (VILLAVICENCIO et al., 2007). Pode-se, por exemplo, mais que dobrar o número de entradas de construções verbo-

partícula (CVP) em um dicionário através da extração dessas em um *corpus* como o BNC (BALDWIN, 2005). São utilizadas para esta tarefa, desde informações puramente estatísticas (como medidas de associação), a informações linguísticas como, por exemplo, filtragens através de informações sintáticas, demonstrando diferentes graus de sucesso (EVERT; KRENN, 2005; BALDWIN, 2005).

Algumas EMPs são fixas e não apresentam variação interna como *ad hoc*, enquanto outras permitem diferentes graus de variabilidade interna, como *touch a nerve* (*touch/find a nerve*) e *spill beans* (*spill several/musical/mountains of beans*). Em termos de semântica, algumas EMPs são mais opacas em seus significados (e.g. *kick the bucket*, que significa *morrer*) enquanto outras possuem um significado mais claro que pode ser inferido das palavras que compõem a expressão (e.g. em *eat up*, a partícula *up* adiciona um sentido de completude ao verbo *eat*). Portanto, devido à heterogeneidade das EMPs, prover métodos apropriados para a identificação automática e tratamento dessas expressões é um desafio real para sistemas de PLN (SAG et al., 2002).

## 2.2 Métodos para identificação de EMPs

Uma variedade de abordagens tem sido proposta para se identificar automaticamente EMPs, diferindo basicamente a qual tipo de EMP e idioma essas abordagens são aplicadas. Embora alguns trabalhos busquem identificar qualquer tipo de EMP (ZHANG et al., 2006; VILLAVICENCIO et al., 2007), devido a heterogeneidade das mesmas, diversos trabalhos visam extrair tipos específicos de expressões, como colocações (PEARCE, 2002), compostos nominais (KELLER; LAPATA, 2003) e CVPs (BALDWIN, 2005; VILLAVICENCIO, 2005; RAMISCH et al., 2008). Alguns dos trabalhos concentram-se em idiomas particulares, como o inglês (PEARCE, 2002; BALDWIN, 2005) e o chinês (PIAO et al., 2006), enquanto outros trabalhos se beneficiam de assimetrias entre as línguas, utilizando informações de uma língua para ajudar a lidar com EMPs em outro idioma. (Villada Moirón; TIEDEMANN, 2006; CASELI et al., 2009).

De forma a determinar se uma dada sequência de palavras é, de fato, uma EMP (*ad hoc* x *the small boy*), certos trabalhos empregam conhecimento linguístico para a tarefa (VILLAVICENCIO, 2005), enquanto outros utilizam métodos estatísticos (PEARCE, 2002; EVERT; KRENN, 2005; ZHANG et al., 2006; VILLAVICENCIO et al., 2007), ou os combinam com alguns tipos de informação linguística como propriedades sintáticas e semânticas (BALDWIN; VILLAVICENCIO, 2002; Van de Cruys; Villada Moirón, 2007) ou, ainda, com técnicas de alinhamento automático de palavras em textos paralelos (Villada Moirón; TIEDEMANN, 2006).

Para o português, a combinação de algumas medidas baseadas em frequências e heurísticas para a identificação de termos para a construção de uma ontologia a partir de textos de domínio específico resultou em uma medida $F_1$ de até 11,51% para bigramas e 8,41% para trigramas (VIEIRA et al., 2009).

### 2.2.1 Medidas de associação

Medidas de associação (MA) tem sido largamente utilizadas na verificação de EMPs pelo fato das mesmas serem meios de baixo custo e de operarem independentes de tipo de EMP e idioma. Como espera-se que as palavras que compõem uma EMP ocorram frequentemente juntas, essas medidas podem, então, indicar que

tal expressão é uma EMP. Desta forma, se um grupo de palavras co-ocorrem com uma frequência significativamente alta em comparação às frequências das palavras individuais, então essas talvez formem uma EMP.

Medidas como Pointwise Mutual Information (PMI), Mutual Information (MI), $\chi^2$, log-likelihood e outras tem sido utilizadas para essa tarefa (PRESS et al., 1992). Algumas dessas medidas parecem fornecer predições mais precisas que outras. Contudo, não há consenso sobre qual medida é melhor para identificar EMPs em geral. Em VILLAVICENCIO et al. (2007), comparou-se algumas dessas medidas (*MI*, *permutation entropy* e $\chi^2$) para a identificação de EMPs independentes de tipo e se concluiu que o método MI pareceu diferenciar EMP de outras expressões, porém o mesmo não é verdade para o método $\chi^2$. Já EVERT; KRENN (2005) observou que a eficácia do processo de identificação de EMPs ao se utilizar uma certa MA depende de fatores como o tipo de EMP buscada, o domínio e tamanho do *corpus* utilizado, bem como a quantidade de dados excluídos ao se adotar um limiar de frequência. Entretanto, VILLAVICENCIO et al. (2007), ao se dicutir a influência da natureza e do tamanho do *corpus*, constatou que essas diferentes medidas tem um alto nível de concordância sobre as EMPs selecionadas, seja em *corpora* cuidadosamente construídos ou em *corpora* mais heterogêneos, como os formados por dados da Web.

Porém, a eficiência destes medidas varia de acordo com as próprias EMPs (por exemplo, tipo e flexibilidade sintática das mesmas) (FAZLY; COOK; STEVENSON, 2009), em características do *corpus* utilizado (como o tamanho e domínio) (VILLAVICENCIO et al., 2007; EVERT; KRENN, 2005) e as listas de referência usadas durante a avaliação (VILLAVICENCIO; CASELI; MACHADO, 2009).

As medidas de associação utilizadas no trabalho são listadas abaixo.

- Pointwise Mutual Information

$$PMI(x,y) = log\frac{p(x,y)}{p(x)p(y)},$$

  onde $p(x)$ e $p(y)$ são as probabilidades de $x$ ou $y$ ocorrerem independentemente e $p(x,y)$ é a probabilidade de $x$ e $y$ ocorrem juntos.

- Mutual Information

$$MI(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right),$$

  onde $p(x)$ e $p(y)$ são as probabilidades de $x$ ou $y$ ocorrerem independentemente e $p(x,y)$ é a probabilidade de $x$ e $y$ ocorrem juntos.

- T-score

$$t = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}},$$

  onde $\bar{x}$ é a média observada, $s^2$ a variância, $N$ é o tamanho da amostra e $\mu$ a média da distribuição.

- Teste $\chi^2$ de Pearson

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

onde $i$ varia entre as linhas de tabelas de contingência construídas com as frequências observadas no *corpus*, $j$ varia entre as colunas, $O_{ij}$ o valor observado na posição (i,j) da tabela de contingência e $E_{ij}$ o valor esperado.

- Coeficiente Dice

$$Dice(x, y) = \frac{2p(x, y)}{(p(x) + p(y))},$$

onde $p(x)$ e $p(y)$ são as probabilidades de $x$ ou $y$ ocorrerem independentemente e $p(x, y)$ é a probabilidade de $x$ e $y$ ocorrerem juntos.

- Teste exato de Fisher

$$\frac{(a + b)!(c + d)!(a + c)!(b + d)!}{n!a!b!c!d!},$$

onde $a$ é o número de ocorrências do bigrama $w_1w_2$, $b$ é o número de ocorrências do bigrama $w_1w_n$ (com $w_n$ diferente de $w_2$), $c$ é o número de ocorrências do bigrama $w_nw_2$ (com $w_n$ diferente de $w_1$) e $d$ é o número de ocorrências do bigrama $w_mw_n$ (com $w_m$ diferente de $w_1$ e $w_n$ diferente de $w_2$).

- Medida de Poisson-Stirling

$$PS = x(\log \frac{x}{y} - 1),$$

onde $x$ é o número de vezes que dois eventos ocorrem juntos e $y$ o valor teórico esperado para essa co-ocorrência.

- Odds Ratio

$$\frac{p_{11}p_{00}}{p_{10}p_{01}},$$

onde $p_{11}$ é a probabilidade que ocorrência do n-grama $w_1w_2$, $p_{00}$, é a probabilidade que ocorrência do n-grama $w_mw_n$ (com $w_m$ diferente de $w_1$ e $w_n$ diferente de $w_2$), $p_{10}$ é a probabilidade que ocorrência do n-grama $w_1w_n$ (com $w_n$ diferente de $w_2$) e $p_{01}$ é a probabilidade que ocorrência do n-grama $w_mw_2$ (com $w_m$ diferente de $w_1$)

### 2.2.2 Alinhamento de textos paralelos

A abordagem baseada em alinhamento utiliza textos paralelos alinhados para identificar candidatos a EMP, isto é, textos acompanhados de suas traduções onde há marcações que identificam as correspondências entre dois textos (CASELI; NUNES, 2005). *Corpora* paralelos podem ser alinhados em relação aos capítulos, às seções, às frases dos textos (alinhamento sentencial), às palavras dos textos (alinhamento lexical), entre outros.

Essa abordagem compara o alinhamento lexical automático das versões em português e inglês do *Corpus* de Pediatria, gerado com o alinhador estatístico de palavras GIZA++. A hipótese é que, quando o alinhador lexical encontra uma sequência de palavras no idioma fonte cujas palavras não podem ser alinhadas individualmente ao idioma destino, essa sequência é considerada como uma candidata a EMP. Por exemplo, a sequência em português "aleitamento materno" — que ocorre 202 vezes no *corpus* utilizado nos experimentos — é uma candidata a EMP pois essas duas palavras são agrupadas e alinhadas 184 vezes com a palavra breastfeeding (um alinhamento 2:1), 8 vezes com a palavra breastfed (um alinhamento 2:1), 2 vezes com "brestfeeding practice"(um alinhamento 2:2). Assim, a abordagem baseada em alinhamento considera como candidatas a EMP, sequências de duas ou mais palavras do idioma fonte que são agrupadas pelo alinhador, independente delas serem alinhadas com uma ou mais palavras no texto alvo.

### 2.2.3 Método híbrido

Além de avaliar os métodos estatístico e baseado em alinhamento separadamente, este trabalho descreve um método híbrido, com a combinação desses para se obter um conjunto de candidatos a EMP mais preciso que aqueles fornecidos pelos métodos individualmente. A abordagem foi proposta em RAMISCH et al. (2009), e pode ser utilizada para auxiliar o trabalho lexicográfico provendo uma lista de candidatos a EMP mais precisa, mantendo recursos lexicais atualizados e, também, para melhorar a qualidade de sistemas de PLN.

Para combinar os dois métodos, um classificador foi construído com o uso do pacote de software `Weka` (WITTEN; FRANK, 1999), que agrega algoritmos provenientes de diferentes abordagens na área da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas.

Os n-gramas resultantes após o processo de filtragem sintática (descrito nos experimentos), anotados com o julgamento das diferentes medidas utilizadas no método estatístico, são fornecidos como entradas para um classificador. Esse classificador busca determinar se cada n-grama é, ou não, uma EMP. A tabela 2.1 mostra alguns exemplos de entradas do conjuntos em português e inglês, com os valores de cada MA e uma informação indicando se a utilização do alinhador léxico considera o n-grama como um possível candidato.

Devido ao alto número de n-gramas fornecidos nos conjuntos de treinamento, as classes são consideradas desbalanceadas, ou seja, existem muito menos casos de uma das classes em relação a outra (CHAWLA; JAPKOWICZ; KOTCZ, 2004). Este tipo de desbalanceamento tende a valorizar classes predominantes e a ignorar classes de menor representação (PHUA; ALAHAKOON; LEE, 2004). Para combinar as diferentes abordagens, foi utilizado um classificador Bayesiano, visto que verificou-

Tabela 2.1: Exemplos de entradas dos conjuntos de treinamento.

| n-grama | Align | Estatístico | | | | | | | | EMP |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Dice | Odds | PMI | PS | t-score | MI | $\chi^2$ | Fisher | |
| abnormal findings | Yes | .03 | 114.1 | 6.74 | 25.70 | 2.62 | 0 | 734.73 | 0 | No |
| adrenal insufficiency | No | .46 | 10376 | 11.6 | 371.9 | 7.28 | .0008 | 160784 | 0 | Yes |
| óxido nítrico | Yes | .95 | 8553397 | 14.5 | 289.3 | 5.66 | .0006 | 733177 | 0 | Yes |
| academia americana | No | .52 | 74302 | 13.3 | 197.4 | 4.9 | .0004 | 244244 | 0 | No |

se que esse é robusto e menos sensível a classes altamente desbalanceadas[1].

Um classificador Bayesiano busca determinar a qual classe um certo dado de entrada pertence. Para tanto, é utilizado o teorema de Bayes (mostrado abaixo) para determinar qual a probabilidade de amostra desconhecida pertencer a cada uma das classes possíveis. Assim, assume-se que essa amostra pertence a classe que apresentar a maior probabilidade. Neste trabalho, o classificador Bayesiano é utilizado de forma a classificar os n-gramas em duas classes, buscando determinar se o mesmo é, ou não, uma EMP.

Teorema de Bayes:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

onde $P(A)$ e $P(B)$ é a probabilidade a priori de $A$ e $B$ respectivamente, $P(A|B)$ é a probabilidade condicional de $A$ dada a ocorrência de $B$, $P(B|A)$ é a probabilidade condicional de $B$ dada a ocorrência de $A$.

[1] O uso de árvores de decisão para a língua inglesa gerou uma única classe, classificando negativamente todos os candidatos.

# 3 MATERIAIS

Neste capítulo serão descritos os recursos lexicais utilizados no trabalho. Características do *corpus* paralelo são mostradas e é descrito o processo de criação das listas de referência utilizadas como *gold standard* no processo de avaliação dos métodos de identificação de EMPs.

## 3.1 Corpus de pediatria

Para os experimentos, foi utilizado o Corpus de Pediatria, um *corpus* paralelo composto de 283 textos extraídos do Jornal de Pediatria[1]. Todos esses textos foram publicados em dois idiomas, o português (totalizando 785.448 palavras) e o inglês (729.923 palavras).

Ao utilizarmos um *corpus* paralelo, podemos aplicar o método baseado em alinhamento mostrado na seção 2.2.2, obtendo proveito das assimetrias existentes entre os idiomas para a identificação de expressões multipalavra. Além disso, o uso de um *corpus* paralelo permite também investigar se a escolha do idioma influencia os resultados obtidos, isto é, se as características de um idioma influenciam os resultados obtidos pelos métodos de identificação de EMPs estudados no trabalho.

Um *corpus* de domínio específico foi escolhido pois a terminologia utilizada nesse tipo de texto consiste, em grande parte, de expressões multipalavra. Além disso, percebe-se que, para textos de domínio específico, o vocabulário é mais controlado. Ao se utilizar um *corpus* mais heterogêno, por exemplo, são usadas expressões como "ataque no coração" e "ataque cardíaco" para se referir a um mesmo termo técnico de área de Medicina ("infarto do miocárdio"). Por fim, uma outra justificativa para o uso de *corpora* técnicos trata-se da menor ocorrência de erros (ortográficos, sintáticos, entre outros) nesse tipo de texto, reduzindo a ocorrência de n-gramas raros que poderiam ser considerados como bons candidatos a EMP.

Para os experimentos, o *corpus* foi submetido a um pré-processamento. Todos os caracteres do *corpus* foram representados em caracteres minúsculos. Assim, busca-se unificar em um único n-grama formas distintas como, por exemplo, "Ad Hoc", "Ad hoc", "ad hoc", elevando a frequência do n-grama "ad hoc". Além disso, o *corpus* foi marcado com o uso de analisadores morfológicos e sintáticos do `Apertium`[2] estendido com entradas descritas em ARMENTANO-OLLER et al. (2006). As informações sintáticas são utilizadas para remover n-gramas das listas de candidatos a EMP das abordagens mostradas na seção 4.

[1] `www.jped.com.br`

[2] `Apertium` é sistema de tradução automática disponível em: `http://www.apertium.org`.

Neste trabalho, para auxiliar o processo de identificação de EMPs, os candidatos foram anotados com etiquetas morfossintáticas. Atribui-se para cada palavra (incluindo números e sinais de pontuação), sua classe gramatical. Esse processo é semelhante ao processo de "tokenização" para linguagens computacionais, embora as etiquetas para linguagens naturais apresentem maior ambiguidade. Ferramentas que realizam esta marcação (etiquetadores ou *taggers*) exercem um papel importante em áreas como o reconhecimento de voz, análise sintática de linguagem natural e recuperação de informações. Tais informações podem ser úteis, por exemplo, em um modelo de linguagem para reconhecimento de voz. Saber que uma palavra é um pronome possessivo ou um pronome pessoal, por exemplo, pode indicar quais palavras possivelmente ocorrerão em sua vizinhança (pronomes possessivos são comumente seguidos por um substantivo, pronomes pessoais por um verbo). Segundo JURAFSKY; MARTIN (2008), conhecer a classe gramatical das palavras de uma frase é importante, pois isso fornece informações significativas sobre uma palavra e suas palavras vizinhas.

Abaixo, é mostrado o resultado obtido com o uso do etiquetador morfossintático do sistema `Apertium` para o seguinte trecho presente no *corpus*: "... constituiu um dos principais fatores de risco para a oferta de líquidos suplementares aos neonatos (18).".

constituiu/constituir<vblex><ifi><p3><sg> um/um<num><m><sp>
dos/de<pr>+o<det><def><m><pl> principais/principal<adj><mf><pl>
fatores/fator<n><m><pl> de/de<pr> **risco**/risco<n><m><sg>
para/para<pr> a/o<det><def><f><sg> oferta/oferta<n><f><sg> de/de<pr>
líquidos/líquido<n><m><pl> suplementares/suplementar<adj><mf><pl>
aos/a<pr>+o<det><def><m><pl> neonatos/neonato<n><m><pl>
(/(<lpar> 18/18<num> )/)<rpar> ./.<sent>

Percebe-se, neste exemplo, que atribuir automaticamente uma classe gramatical a cada palavra não é trivial. Observa-se que o termo realçado "risco" é ambíguo. Isto é, há mais de uma classificação gramatical possível. A palavra pode ser um substantivo (como em *o risco de morte*) ou um verbo (*eu risco a parede*). O objetivo dos etiquetadores é, justamente, resolver esse tipo de ambiguidade.

Neste trabalho, a informação morfossintática é usada para filtrar candidatos que em geral não são EMPs, tornando a lista de candidatos dos métodos de identificação de EMPs mais precisas.

## 3.2 Listas de referência

O processo de avaliação automático utiliza o Dicionário de Pediatria[3], um recurso lexical de domínio específico construído de forma semiautomática a partir do *corpus* de Pediatria com o objetivo de fornecer suporte a estudos de tradução[4]. As expressões em português do dicionário foram geradas através da extração de todos os n-gramas (com $n$ entre 2 e 4) do *corpus* que ocorreram ao menos cinco vezes. Após, os candidatos foram anotados com suas etiquetas morfossintáticas e foram removidos os candidatos iniciados por artigo+substantivo e começando ou terminando

[3]Disponível em `http://www6.ufrgs.br/textquim/Dicionarios/DicPed/`
[4]Produzido pelo TEXTQUIM/TERMISUL: `http://www.ufrgs.br/textquim`

por verbos que, em observações empíricas, em geral, não correspondem a EMPs. Neste processo todos os bigramas válidos contidos em trigramas foram adicionados ao dicionário, visto que esses haviam sido retirados durante a construção original do dicionário (LOPES et al., 2009). A construção do dicionário em inglês seguiu um processo similar e foi baseado nas traduções correspondentes de todas as expressões geradas para o português.

As versões finais das listas de referência possuem 2.150 termos em português e 883 termos em inglês. Em virtude do menor número de entradas no dicionário em inglês, para a avaliação dos candidatos também foram usadas expressões provenientes de um dicionário geral do inglês, o *Cambridge International Dictionary of Idioms* (CIDE) (CAMBRIDGE, 1994). As EMPs do Dicionário de Pediatria são consideradas específicas de domínio e as do CIDE genéricas, o número de entradas de cada um são mostradas na tabela 3.1. Exemplos de entradas das listas de referências são mostradas na tabela 3.2.

Tabela 3.1: Número de entradas em cada lista de referência

| | Específico | Genérico | Total |
|---|---|---|---|
| Português | 2150 | — | 2150 |
| Inglês | 883 | 1382 | 2190 |

Tabela 3.2: Exemplos de entradas de cada lista de referência

| **Português** | **Inglês** |
|---|---|
| abertura traqueal | absence of t |
| abordagem diagnóstica | accidental extubation |
| abordagem terapêutica | acetyl salicylic acid |
| abscesso mamário | action of insulin |
| absorção de cálcio | acute asthma crises |
| absorção de vitamina | acute bronchiolitis |
| absorção intestinal | acute leukemia |
| acalasia de esôfago | acute myeloid leukemia |
| ação da insulina | acute otitis media |
| ácaros da poeira | acute pyelonephritis |

# 4 EXPERIMENTOS

Nesta seção, são descritos os experimentos realizados para se avaliar a eficácia das abordagens para a identificação de EMPs apresentadas na seção 2. Os resultados são analisados e uma comparação dos métodos é feita.

## 4.1 Avaliação de sistemas de recuperação de informações

Como forma de avaliar os resultados obtidos pelos diferentes métodos, são calculadas as medidas de **abrangência**, **precisão** e $F_1$ (obtida através da **medida F**):

- A medida de *abrangência* (também chamada de revocação, do inglês *recall*) indica o quanto de informação relevante foi extraída do texto ou *corpus* e é definida como segue:

$$\text{Abrangência} = \frac{\text{número de expressões corretas fornecidas pelo sistema}}{\text{número de expressões corretas no corpus}}$$

- A *precisão* indica o quanto de informação retornada pelo sistema é correta. Tal medida é definida como segue:

$$\text{Precisão} = \frac{\text{número de expressões corretas fornecidas pelo sistema}}{\text{número de expressões fornecidas pelo sistema}}$$

- A medida $F$, que combina as medidas anteriores. Essa medida é definida como segue:

$$F = \frac{(1 + \beta^2) \times \text{precisão} \times \text{abrangência}}{\beta^2 \times \text{precisão} + \text{abrangência}}$$

Observa-se que as medidas de abrangência e precisão são complementares uma a outra. Um sistema conservador que busca uma alta precisão apresentará invariavelmente um baixo valor de abrangência. Similarmente, um sistema que busca atingir uma alta cobertura apresentará diversas respostas incorretas, o que ocasionará um baixo valor de precisão. Por isso, é também usada a medida F, que combina precisão e abrangência. Quando $\beta$ é igual a 1, obtém-se a medida $F_1$, que equivale à média harmônica das medidas de precisão e abrangência:

$$F_1 = \frac{2 \times \text{precisão} \times \text{abrangência}}{\text{precisão} + \text{abrangência}}$$

## 4.2 Geração de candidatos para os métodos estatístico e baseado em alinhamento

Esta seção mostra o processo realizado para a geração de candidatos para os métodos estatístico e baseado em alinhamento lexical, bem como a comparação do número de candidatos extraídos por cada abordagem.

Nos experimentos, são extraídas apenas sequências de palavras contendo duas ou três palavras (bigramas e trigramas, respectivamente). Essa decisão foi tomada pois grande parte das expressões presentes nas listas de referência usadas durante a avaliação possuem duas ou três palavras. Além disso, aumentar o número de palavras por expressão tornaria os experimentos mais custosos computacionalmente, demandando mais recursos de processamento e armazenamento. Os experimentos realizados foram desenvolvidos de forma a permitir que esses sejam estendidos para outros idiomas e tamanhos de n-gramas. Porém, como não seria possível avaliar os resultados, foram focados apenas bigramas e trigramas.

### 4.2.1 Geração de candidatos para método estatístico

Para os dois conjuntos de textos do *corpus* (português e inglês), as expressões candidatas foram geradas a partir de todos os bigramas e trigramas contendo palavras/numerais/sinais de pontuação do Corpus de Pediatria. Como exemplo, os bigramas e trigramas gerados para a frase *"a mãe foi transferida em caráter de emergência para o nosso hospital."* são mostrados na tabela 4.1.

A abordagem estatística não combina os textos em português e inglês, gerando listas de candidatos independentes, pois se considera separadamente os conjuntos de textos nos diferentes idiomas. Dessa forma, foram gerados 244.420 bigramas e 513.494 trigramas para o português e 230.130 bigramas e 492.154 trigramas para o *corpus* em inglês. Visto que nenhuma das expressões presentes nas listas de referência continham sinais de pontuação e numerais, os n-gramas que possuíam esses caracteres foram removidos das listas de candidatos. Ao final dessa primeira filtragem, foram gerados um total de 185.377 bigramas e 326.745 trigramas para o português e 175.881 bigramas e 316.384 trigramas candidatos a EMP para a língua inglesa.

Tabela 4.1: Bigramas e trigramas gerados para a frase *a mãe foi transferida em caráter de emergência para o nosso hospital.*

| Bigramas | Trigramas |
|---|---|
| a mãe | a mãe foi |
| mãe foi | mãe foi transferida |
| foi transferida | foi transferida em |
| transferida em | transferida em caráter |
| em caráter | em caráter de |
| caráter de | caráter de emergência |
| de emergência | de emergência para |
| emergência para | emergência para o |
| para o | para o nosso |
| o nosso | o nosso hospital |
| nosso hospital | nosso hospital . |
| hospital . | |

Tabela 4.2: N-gramas mais frequentes no *corpus*

| Português | Frequência | Inglês | Frequência |
|---|---|---|---|
| para a | 1190 | of the | 6338 |
| e a | 1132 | in the | 4053 |
| que a | 964 | to the | 1708 |
| para o | 847 | and the | 1403 |
| com a | 834 | for the | 1334 |
| com o | 829 | with the | 1187 |
| uso de | 788 | it is | 1090 |
| e o | 785 | by the | 1056 |
| et al | 763 | use of | 989 |
| em crianças | 758 | on the | 984 |

Percebe-se que apenas escolher os n-gramas mais frequentes como bons candidatos a EMP não é interessante, o que pode ser observado na tabela 4.2. Essa tabela mostra os n-gramas mais frequentes no *corpus* e suas frequências. A grande ocorrência de artigos, conjunções e verbos auxiliares no *corpus* gera candidatos para o inglês como *of the*, *and the* e *there are*. Um comportamento semelhante é observado para os candidatos em português.

De forma a melhorar a precisão dos resultados, deve-se retirar o máximo possível de n-gramas ruidosos da lista de candidatos a EMP. Assim, foram descartados todos aqueles n-gramas que (a) seguiam certos padrões de etiquetas morfossintáticas ou (b) que ocorreram menos de cinco vezes. Esse limiar de frequência foi utilizado de forma a excluir candidatos que apresentaram uma frequência muito baixa, visto que esses introduzem ruído para os métodos utilizados. Além disso, a partir de uma única ocorrência de um candidato, não há subsídios suficientes para diferenciar entre ruído (por exemplo, um erro de digitação), e uma EMP rara.

Neste trabalho, o processo de etiquetação morfossintática é utilizado como forma de descartar n-gramas que, segundo observações empíricas, em geral, não são EMPs das listas de candidatos. Busca-se, assim, limpar a lista de candidatos tornando-a mais precisa. Os padrões de filtragem utilizados são os definidos por CASELI et al. (2009) e são reproduzidos na tabela 4.3, juntamente com exemplos de sequências de palavras filtradas. É importante verificar que este processo pode ocasionar a remoção de falsos negativos, como as EMPs *his Majesty* e *My God*. Os números de candidatos restantes após o processo de filtragem são mostrados neste capítulo, na tabela 4.4. No trabalho de CASELI et al. (2009), os filtros utilizados são definidos apenas para o inglês. De forma a realizar a filtragem para os candidatos em português, foram definidas regras equivalentes às mostradas na tabela 4.3. Como exemplo, para a regra mostrada na sexta linha da tabela, a filtragem utilizada para o português inclui palavras como “são”, “é”, “era”, “eram”.

### 4.2.2 Geração de candidatos para método baseado em alinhamento lexical

Esta abordagem é baseada no alinhamento lexical automático das versões em português e inglês do Corpus de Pediatria e foi gerado com o alinhador estatístico

Tabela 4.3: Regras utilizadas para filtragem de candidatos

| Primeiro termo do n-grama | Exemplos de candidatos filtrados |
|---|---|
| artigo | a detector, a cure, an increase, the american, the atmospheric institute |
| verbo auxiliar | does exist, did not, did you, had become, will be, will gain, would allow |
| pronome | he called, he argues, their children, his life, these, are, this spirit |
| advérbio | widely studied, publicly stored, not yet, since then, under suspicion |
| conjunção | as smoke cover, or produces, as in workers, and yet, and hence |
| are, is, was, were (verbo ser/to be) | are already, are a result, is to, were able, was formed |
| that, what, when, which, who, why | that are, that varies, what was, why do, which lasts, who responds |
| from, to, of | from them, from Bahia, to build, to the, of cell, of our, of this |

de palavras GIZA++[1]. A abordagem baseada em alinhamento foi proposta em CASELI et al. (2009) e considera como candidatas a EMP, as sequências de duas ou mais palavras que são agrupadas pelo alinhador, sem considerar se essas são alinhadas com uma ou mais palavras no texto destino.

Após o processo de alinhamento lexical e geração de candidatos, esses foram filtrados através de suas características morfossintáticas de forma semelhante ao método estatístico.

Dado que o método baseado em alinhamento busca sequências de palavras que são frequentemente agrupadas juntas no processo de alinhamento, esse método resulta em uma lista menor de candidatos a EMP (pois nem todas as sequências frequentes de palavras são consideradas), mas com maior precisão nos candidatos gerados.

### 4.2.3 Comparação de candidatos gerados para os métodos estatístico e baseado em alinhamento

A tabela 4.4 mostra o número candidatos originais extraídos para cada idioma antes e após as filtragens, Para a abordagem baseada em alinhamento, além dos bigramas e trigramas, é mostrado também o número de candidatos onde o número de palavras é maior que 3, porém, esses não são considerados nos experimentos executados.

Por um lado, ambos os idiomas possuem aproximadamente o mesmo número de candidatos para cada abordagem, observa-se que esse o processo de filtragem reduz consideravelmente as listas de candidatos. Por outro lado, como um limiar de frequência mais alto foi aplicado para as medidas estatísticas (5 ocorrências, contra 1 do método de alinhamento), há um maior número de candidatos para o método baseado em alinhamento. A última seção da tabela 4.4 indica que ambas

[1]Agradecemos a Helena Caseli por disponibilizar os dados de alinhamento lexical para o Corpus de Pediatria. O alinhador usado é descrito em CASELI (2007).

as abordagens são essencialmente diferentes em relação aos candidatos extraídos: menos de 15% dos candidatos extraídos pelo método baseado em alinhamento são também capturados pelo método estatístico e vice-versa.

Tabela 4.4: Número de candidatos gerados para cada método, separados por idioma e tamanho do n-grama

| **Candidatos sem filtragem** | | | | | **Candidatos após filtragem** | | | |
|---|---|---|---|---|---|---|---|---|
| | n-gramas | | | | Estatístico (stat) | | | |
| | $n=2$ | $n=3$ | $n>3$ | Total | $n=2$ | $n=3$ | $n>3$ | Total |
| pt | 244420 | 513494 | — | 757914 | 11290 | 4553 | — | 15843 |
| en | 230130 | 492154 | — | 722284 | 10311 | 4526 | — | 14837 |
| | Baseado em alinhamento | | | | Baseado em alinhamento (align) | | | |
| | $n=2$ | $n=3$ | $n>3$ | Total | $n=2$ | $n=3$ | $n>3$ | Total |
| pt | 15333 | 7373 | 11571 | 34277 | 12154 | 5518 | 7117 | 24789 |
| en | 16345 | 7469 | 12649 | 36463 | 12222 | 5154 | 6384 | 23760 |
| | stat ∩ align | | | | | | | |
| | $n=2$ | | $n=3$ | | $n>3$ | | | Total |
| pt | 1376 | | 134 | | — | | | 1510 |
| en | 1921 | | 109 | | — | | | 2030 |

## 4.3 Identificação de EMPs utilizando medidas de associação

Todas os n-gramas resultantes após os processos de geração e filtragem mostrados na seção 4.2.1 foram, então, avaliados utilizando as medidas Pointwise Mutual Information, Mutual Information, t-score, Teste $\chi^2$ de Pearson, coeficiente Dice, teste exato de Fisher, medida de Poisson-Stirling e Odds Ratio (PRESS et al., 1992). Todas as medidas são implementadas pelo Ngram Statistics Package (BANERJEE; PEDERSEN, 2003). Um valor de medida alto para um n-grama indica uma alta confiança de que as palavras do n-grama sejam dependentes, indicando que esse n-grama é um bom candidato a EMP. Como são aplicadas diversas medidas a cada n-grama, a classificação final desse n-grama é dada pela média das posições assumidas por cada n-grama em cada medida. Assim, n-gramas que apresentem uma boa classificação em diversas métricas são considerados bons candidatos a EMP.

Como exemplo, a tabela 4.5 mostra as cinco expressões mais bem classificadas e as cinco com menor pontuação retornadas pela média dos métodos PMI e MI sem a utilização de filtragens. Embora alguns dos resultados sejam bons, especialmente os candidatos do topo, ainda se percebe ruído nos resultados, como a expressão *jogar video game*. Percebe-se, também, que os candidatos com pior classificação são, justamente, n-gramas ruidosos, iniciados por palavras muito frequentes na língua portuguesa (e.g., "e", "do", "que"). As medidas utilizadas não identificam apenas expressões da área de pediatria (*pneumocystis carinii*, *artrites idiopáticas juvenis*) mas também outras possíveis EMP como *vigilância sanitária* e *estados unidos*.

Com o objetivo de se avaliar os resultados obtidos, são mostrados valores de precisão, abrangência e $F_1$ na tabela 4.6. Essa tabela considera a lista de candidatos

Tabela 4.5: 5 melhores e piores candidatos para trigramas utilizando a média dos resultados de PMI e MI

| **Expressões obtidas** |
|---|
| online mendelian inheritance |
| beta technology incorporated |
| lange beta technology |
| oxido nítrico inalatório |
| jogar video game |
| ... |
| e um de |
| e a do |
| se que de |
| e a da |
| e de não |

após realizadas as filtragens por frequência e por sequências de etiquetas morfossintáticas. Os bigramas e trigramas de cada idioma são agrupados, visto que as listas de referência utilizadas não diferenciam o tamanho do n-grama. Verifica-se um alto valor de abrangência (principalmente para o português), ou seja, grande parte das EMPs presentes nas listas de referência foram identificadas. Porém, um grande número de expressões ruidosas foram sugeridas como candidatas a EMP, justificando os baixos valores de precisão obtidos.

Tabela 4.6: Verdadeiros positivos (VP), precisão, abrangência e $F_1$ para método estatístico

| | VP | Precisão | Abrangência | $F_1$ |
|---|---|---|---|---|
| pt_spec | 1852 | 11.69% | 86.14% | 20.59% |
| en_spec | 601 | 4.05% | 68.06% | 7.65% |
| en_spec+gen | 774 | 5.22% | 35.34% | 9.10% |

A Tabela 4.7 mostra o número de candidatos em cada lista de referência considerando os dicionários de pediatria em português e inglês (pt_spec e en_spec) e aquelas no dicionário genérico para a língua inglesa (en_spec+gen). A diferença nos resultados para os idiomas pode ser explicada devido às diferenças na cobertura dos *gold standards* (o glossário em português contém um número maior de entradas). Assim, os resultados para a língua portuguesa utilizando o método estatístico identifica 86,14% das EMPs do *corpus* com uma precisão de 11,69%. Já para o inglês, apenas 35% das instâncias do *gold standard* são identificadas com uma precisão de 5%, utilizando tanto expressões de domínio específico e expressões genéricas. Como a versão estendida do *gold standard* em inglês melhora a $F_1$, adota-se essa versão nas próximas avaliações.

## 4.4 Identificação de EMPs utilizando alinhamento de textos paralelos

Em comparação com a abordagem estatística, o método baseado em alinhamento apresenta uma performance mais baixa. Isso ocorre pois o número de candidatos considerados é maior (conforme a tabela 4.4). Um limiar de frequência mais alto poderia ser utilizado de forma a reduzir o número de candidatos a EMP. Porém, filtros mais restritivos não foram aplicados pois desejou-se investigar como a combinação desses métodos pode remover n-gramas ruidosos das listas de candidatos.

Tabela 4.7: Verdadeiros positivos (VP), precisão, abrangência e $F_1$ do método baseado em alinhamento

| | VP | Precisão | Abrangência | $F_1$ |
|---|---|---|---|---|
| pt_spec | 240 | 0.97% | 11.16% | 1.78% |
| en_spec | 84 | 0.35% | 9.51% | 0.68% |
| en_spec+gen | 224 | 0.94% | 10.23% | 1.72% |

## 4.5 Método híbrido para a identificação de EMPs

Após avaliar as abordagens de forma independente, foi avaliado o método híbrido, que combina as duas abordagens mostradas acima (conforme descrito na seção 2.2.3). Com o objetivo de avaliar a contribuição particular de cada métodos para os resultados, considerou-se quatro conjuntos de atributos:

**subAM** Um subconjunto das medidas de associação (PMI, PS e MI), medidas que não envolvem a construção de tabelas de contingência e podem ser aplicadas diretamente a n-gramas de tamanho arbitrário

**subAM+align** Combinação do subconjunto subAM e abordagem baseada em alinhamento

**allAM** Inclui subAM para bigramas e trigramas e outras MAs apenas para bigramas

**allAM+align** A combinação de todas as MAs e do método baseado em alinhamento

Os resultados ao fornecer os conjuntos de treinamento de cada língua para uma rede Bayesiana, utilizando 10-fold cross validation são mostrados na tabela 4.8[2]. Para ambos os idiomas, o modelo híbrido é capaz de gerar candidatos muito melhores que as listas propostas pelos métodos individualmente. Por exemplo, a rede Bayesiana retorna resultados onde o $F_1$ é de aproximadamente 50% contra 20,59% e 1,79% para os métodos estatístico e baseado em alinhamento, respectivamente.

Em termos de atributos individuais, a adição da informação de alinhamento geralmente melhora a abrangência dos resultados. Porém isso provoca a redução da precisão, adicionando ruído aos resultados e reduzindo a performance geral. Uma

[2]Agradecemos a Carlos Ramisch por disponibilizar os resultados referentes à avaliação do método híbrido.

Tabela 4.8: Classificador Bayesiano para diferentes conjuntos de atributos, em português e inglês

| **Português** | VP | Precisão | Abrangência | $F_1$ |
|---|---|---|---|---|
| subAM | 1102 | 48.29% | 51.26% | 49.73% |
| subAM + align | 1103 | 47.98% | 51.30% | 49.58% |
| allAM | 1100 | 43.51% | 51.16% | 47.03% |
| allAM + align | 1084 | 43.41% | 50.42% | 46.65% |
| **Inglês** | VP | Precisão | Abrangência | $F_1$ |
| subAM | 0 | — | — | — |
| subAM + align | 62 | 16.49% | 2.88% | 4.91% |
| allAM | 464 | 19.74% | 21.58% | 20.62% |
| allAM + align | 465 | 19.68% | 21.63% | 20.61% |

exceção é o conjunto subAM para o inglês, onde a rede Bayesiana, utilizando apenas o subAM não retorna qualquer resultado correto. Porém, ao se adicionar a informação de alinhamento, obteve-se informação suficiente para que a rede apresente uma performance de 4,91%. Isso sugere que a informação de alinhamento pode ajudar a adicionar robustez ao processo.

De forma a avaliar a contribuição do atributo de alinhamento, construiu-se uma árvore de decisão com os mesmos conjuntos de treinamento para a rede Bayesiana. Árvores de decisão buscam classificar os dados do conjunto de treinamento nas diferentes classes possíveis através da melhor combinação possível dos atributos de entrada. Por exemplo, considerando as entradas da tabela 2.1, o atributo *t-score*, quando comparado com o valor 5.0, identifica os n-gramas que são EMPs (valor maior que 5.0) e quais as expressões não são EMPs (*t-score* menor que 5.0). São considerados como melhores atributos aqueles que são capazes de separar a maior quantidade de entradas em suas classes respectivas. Inicialmente, comparações envolvendo esses atributos são realizadas e, posteriormente, outros atributos são utilizados para refinar o processo de classificação.

Assim, analisou-se a profundidade da primeira ocorrência do atributo de alinhamento nas árvores resultantes. Observou-se que as medidas PMI, PS e Dice parecem apresentar melhores indicativos para a identificação de EMPs, pois o atributo de alinhamento é somente usado após essas medidas. A adição de outras MAs parece ter efeito similar, tornando mais robusta a tarefa. Observa-se, também, na tabela 4.8, que essa adição provê maior contribuição para a língua inglesa, enquanto que, para o português, a adição dessas medidas introduz ruído aos resultados. Essas resultados estão de acordo com os conclusões de EVERT; KRENN (2005), as medidas de associação são altamente dependentes do tipo de EMP e idioma do *corpus*.

Algumas das MAs usadas como atributos são baseadas em tabelas de contingência e não podem ser diretamente aplicadas a trigramas[3]. Dessa forma, essas medidas são representadas com um valor desconhecido ("?") para trigramas. Com o objetivo de verificar em mais detalhes se esses valores afetavam os resultados obtidos, foi realizada uma segunda avaliação onde foram analisados somente os candidatos

[3]O NSP, por exemplo, não implementa essas medidas para trigramas.

Tabela 4.9: Performance do classificador para diferentes conjuntos de atributos e idiomas utilizando apenas bigramas.

| **Português** | VP | Precisão | Abrangência | $F_1$ |
|---|---|---|---|---|
| subAM | 1021 | 49.11% | 71.90% | 58.36% |
| subAM + align | 1026 | 45.72% | 72.25% | 56.00% |
| allAM | 1113 | 43.04% | 78.38% | 55.57% |
| allAM + align | 1113 | 42.04% | 78.38% | 55.57% |
| **Inglês** | VP | Precisão | Abrangência | $F_1$ |
| subAM | 228 | 32.66% | 16.06% | 21.53% |
| subAM + align | 267 | 26.67% | 18.80% | 22.06% |
| allAM | 459 | 19.94% | 32.32% | 24.66% |
| allAM + align | 459 | 19.94% | 32.32% | 24.66% |

que não possuíam valores desconhecidos, isto é, os bigramas.

A performance do classificador construído com o conjunto de bigramas é mostrado na tabela 4.9[4]. Os resultados comprovam que, em alguns casos, esses atributos extras adicionam informação suficiente para melhorar a performance do classificador Bayesiano. Isso é facilmente percebido ao se comparar os conjuntos subAM e allAM para o inglês, onde a $F_1$ aumenta de 16,06% para 32,32%. A diferença dos resultados obtidos ao se considerar apenas os bigramas sugere que os métodos propuseram um conjunto de candidatos mais preciso para bigramas, mas esses não parecem tão efetivos para trigramas. Investigação futura será realizada para avaliar apropriadamente que fatores que causam a baixa performance para trigramas.

[4]Para este experimento, foram considerados apenas os bigramas dos *gold standards*

# 5 CONCLUSÕES E TRABALHOS FUTUROS

Expressões multipalavra formam um conjunto heterogêneo e complexo, tornando a identificação de todas as suas ocorrências um desafio. Porém, devido ao seu importante papel na linguagem, observa-se uma grande necessidade de que essas expressões sejam identificadas de forma apropriada.

Nesse trabalho, foi investigada a tarefa de identificação de EMPs em domínios técnicos através de diferentes abordagens. Os métodos investigados foram o método baseado em informações estatísticas, baseado em alinhamento lexical e uma abordagem híbrida, que combina os dois métodos anteriores. Para os experimentos, foi utilizado um *corpus* paralelo do domínio de Pediatria.

A abordagem estatística mostrou resultados com alto valor de abrangência e baixo valor de precisão, isto é, grande parte das expressões do *gold standard* foram identificadas. Porém, tal método considerou como EMP muitos n-gramas que não eram, efetivamente, expressões multipalavra. Já a abordagem baseada em alinhamento lexical apresentou um baixo desempenho em comparação ao obtido pelo método estatístico. Isso deve-se ao fato de não ter sido utilizado um filtro de frequência com um limiar mais alto, como ocorre na abordagem estatística.

Também foi apresentada uma combinação dos métodos de forma a produzir um conjunto de candidatos a EMP com maior precisão que os métodos estatísticos e apresentando um valor de abrangência maior que os apresentados por métodos baseados em alinhamento. O uso da informação de alinhamento, bem como um conjunto maior de medidas de associação, adicionaram robustez ao processo de identificação de EMPs, provendo informação suficiente para o classificador.

Como trabalhos futuros, pretende-se conduzir experimentos de forma a identificar quais fatores determinam a influência do atributo de alinhamento no resultado final, visto que esta informação pode tanto introduzir ruído no resultado como melhorar a performance do classificador de acordo com o idioma e tamanho do n-grama. Além disso, planeja-se investigar a influência do domínio do *corpus* na performance dos métodos apresentados, verificando se a extração de EMP de domínio específico é realizada de forma mais fácil que expressões genéricas.

Por fim, experimentos realizados indicam que, embora a filtragem de termos utilizando informações sintáticas retire efetivamente sequências de palavras ruidosas das listas de candidatos, essa pode ser aprimorada. Dessa forma, novas regras de filtragem podem ser definidas a fim de se obter resultados mais precisos sem interferir o valor de abrangência obtido. Entre as regras que poderiam ser incluídas encontram-se, por exemplo, remover expressões onde o último termo é uma preposição ou conjunção (*absoluto de* e *agitação ou*, respectivamente). Investigações sobre quais os de etiquetas sintáticas podem ser filtrados serão realizadas em traba-

lhos futuros.

# REFERÊNCIAS

ARMENTANO-OLLER, C.; CARRASCO, R. C.; CORBÍ-BELLOT, A. M.; FORCADA, M. L.; GINESTÍ-ROSELL, M.; ORTIZ-ROJAS, S.; PÉREZ-ORTIZ, J. A.; RAMÍREZ-SÁNCHEZ, G.; SÁNCHEZ-MARTÍNEZ, F.; SCALCO, M. A. Open-source Portuguese-Spanish machine translation. In: Proceedings of the VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR-2006), 2006, Itatiaia-RJ, Brazil. **Anais...** [S.l.: s.n.], 2006. p.50–59.

BALDWIN, T. The deep lexical acquisition of English verb-particles. **Computer Speech and Language, Special Issue on Multiword Expressions**, [S.l.], v.19, n.4, p.398–414, 2005.

BALDWIN, T.; BENDER, E. M.; FLICKINGER, D.; KIM, A.; OEPEN, S. Road-testing the English Resource Grammar over the British National Corpus. In: FOURTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2004), 2004, Lisbon, Portugal. **Anais...** [S.l.: s.n.], 2004.

BALDWIN, T.; VILLAVICENCIO, A. Extracting the Unextractable: a case study on verb-particles. In: CONFERENCE ON NATURAL LANGUAGE LEARNING (CONLL-2002), 6., 2002, Taipei, Taiwan. **Proceedings...** [S.l.: s.n.], 2002.

BANERJEE, S.; PEDERSEN, T. The Design, Implementation and Use of the Ngram Statistics Package. In: IN PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS, 2003. **Anais...** [S.l.: s.n.], 2003. p.370–381.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. **Grammar of Spoken and Written English**. Harlow: Longman, 1999.

CAMBRIDGE. **Cambridge International Dictionary of English**. [S.l.]: Cambridge University Press, 1994.

CASELI, H. M. **Indução de léxicos bilíngües e regras para a tradução automática**. 2007. Tese (Doutorado em Ciência da Computação) — Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP). 158 p.

CASELI, H. M.; NUNES, M. G. V. Alinhamento Sentencial e Lexical de Córpus Paralelos: recursos para a tradução automática. **Estudos Lingüísticos**, [S.l.], v.34, p.356–361, 2005.

CASELI, H. M.; RAMISCH, C.; NUNES, M. G. V.; VILLAVICENCIO, A. Alignment-based extraction of multiword expressions. **Language Resources and Evaluation**, [S.l.], to appear 2009.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Editorial: special issue on learning from imbalanced data sets. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.6, n.1, p.1–6, 2004.

EVERT, S.; KRENN, B. Using small random samples for the manual evaluation of statistical association measures. **Computer Speech and Language**, [S.l.], v.19, n.4, p.450–466, 2005.

FAZLY, A.; COOK, P.; STEVENSON, S. Unsupervised Type and Token Identification of Idiomatic Expressions. **Computational Linguistics**, [S.l.], v.35, n.1, p.61–103, 2009.

FAZLY, A.; STEVENSON, S. Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, 2007, Prague. **Anais...** [S.l.: s.n.], 2007. p.9–16.

JACKENDOFF, R. Twistin' the night away. **Language**, [S.l.], v.73, p.534–59, 1997.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)**. 2.ed. [S.l.]: Prentice Hall, 2008.

KELLER, F.; LAPATA, M. Using the Web to Obtain Frequencies for Unseen Bigrams. **Computational Linguistics**, [S.l.], v.29, n.3, p.459–484, 2003.

LOPES, L.; VIEIRA, R.; FINATTO, M. J.; ZANETTE, A.; MARTINS, D.; JR., L. C. R. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. **Reciis - Revista Eletrônica de Comunicação Informação & Inovação em Saúde**, [S.l.], v.3, p.76–88, 2009.

PEARCE, D. A Comparative Evaluation of Collocation Extraction Techniques. In: THIRD INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 2002, Las Palmas, Canary Islands, Spain. **Anais...** [S.l.: s.n.], 2002.

PHUA, C.; ALAHAKOON, D.; LEE, V. Minority report in fraud detection: classification of skewed data. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.6, n.1, p.50–59, 2004.

PIAO, S. S. L.; SUN, G.; RAYSON, P.; YUAN, Q. Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool. In: Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), 2006, Trento, Italy. **Anais...** [S.l.: s.n.], 2006. p.17–24.

PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. **Numerical Recipes in C**: the art of scientific computing. second edition. [S.l.]: Cambridge University Press, 1992.

RAMISCH, C.; CASELI, H. M.; VILLAVICENCIO, A.; MACHADO, A.; FINATTO, M. J. A Hybrid A Hybrid Approach for Multiword Expression Identification. In: **Proceedings of the 9th International Workshop on Computational Processing of Written and Spoken Portuguese, (PROPOR 2006)**. [S.l.: s.n.], 2009.

RAMISCH, C.; VILLAVICENCIO, A.; MOURA, L.; IDIART, M. Picking them up and Figuring them out: verb-particle constructions, noise and idiomaticity. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING (CONLL 2008), 12., 2008. **Anais...** [S.l.: s.n.], 2008. p.49–56.

SAG, I. A.; BALDWIN, T.; BOND, F.; COPESTAKE, A.; FLICKINGER, D. Multiword Expressions: a pain in the neck for nlp. In: THIRD INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING (CICLING-2002), 2002, London, UK. **Proceedings...** Springer-Verlag, 2002. p.1–15. ((Lecture Notes in Computer Science), v.2276).

Van de Cruys, T.; Villada Moirón, B. Semantics-based Multiword Expression Extraction. In: Proceedings of the Workshop on A Broader Prespective on Multiword Expressions, 2007, Prague. **Anais...** [S.l.: s.n.], 2007. p.25–32.

VIEIRA, R.; FINATTO, M. J.; MARTINS, D.; ZANETTE, A.; JR, L. C. R. Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. **Reciis - Revista Eletronica de Comunicação Informação e Inovação em Saúde**, [S.l.], v.3, p.76–88, 2009.

Villada Moirón, B.; TIEDEMANN, J. Identifying idiomatic expressions using automatic word-alignment. In: Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006), 2006, Trento, Italy. **Anais...** [S.l.: s.n.], 2006. p.33–40.

VILLAVICENCIO, A. The Availability of Verb-Particle Constructions in Lexical Resources: how much is enough? **Journal of Computer Speech and Language Processing**, [S.l.], v.19, 2005.

VILLAVICENCIO, A.; CASELI, H. M.; MACHADO, A. Identification of Multiword Expressions in Technical Domains: investigating statistical and alignment-based approaches. In: Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology, 2009, São Carlos, SP. **Anais...** [S.l.: s.n.], 2009.

VILLAVICENCIO, A.; KORDONI, V.; ZHANG, Y.; IDIART, M.; RAMISCH, C. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, Prague. **Anais...** [S.l.: s.n.], 2007. p.1034–1043.

WITTEN, I. H.; FRANK, E. **Data Mining**: practical machine learning tools and techniques with java implementations (the morgan kaufmann series in data management systems). 1st.ed. [S.l.]: Morgan Kaufmann, 1999.

ZHANG, Y.; KORDONI, V.; VILLAVICENCIO, A.; IDIART, M. Automated Multiword Expression Prediction for Grammar Engineering. In: WORKSHOP ON MULTIWORD EXPRESSIONS: IDENTIFYING AND EXPLOITING UNDERLYING PROPERTIES, 2006, Sydney, Australia. **Proceedings...** Association for Computational Linguistics, 2006. p.36–44.