Combinatórias Léxicas Especializadas: extração com o uso do `mwetoolkit`

Paulo Guilherme Pilotti Duarte (PIBIC-FAPERGS) – paulopilotti@me.com
Orientadoras: Cleci Regina Bevilacqua – cleci.bevilacqua@ufrgs.br
Coorientadora: Anna Maria Becker Maciel – annamb.maciel@gmail.com

Contextualização: ProjeCOM Legis - coletar as expressões multivocabulares prototípicas da legislação ambiental brasileira e seus equivalentes na legislação do meio ambiente dos países do Mercosul (Argentina, Uruguai e Paraguai), Alemanha, Estados Unidos, França e Itália.

Objetivo do trabalho: Descrever as etapas da aplicação do extrator de expressões multivocabulares (EMs): `mwetoolkit` em um *corpus* de uma área especializada composto por textos legislativos em português brasileiro.

***Corpora* utilizados:**

- De domínio (estudo): *corpus* etiquetado morfossintaticamente da legislação ambiental brasileira (textos legislativos em vigência de 1934 a 2010) da Base Legis do projeto TERMISUL, UFRGS, contendo aproximadamente 600 mil *tokens*.
- De referência: *corpus* da Folha de São Paulo (PLNBRFULL) etiquetado morfossintaticamente, contendo 100 milhões de palavras.

Etapas (ver figura):

- **Indexação (1):** criação de um índice baseado em uma estrutura de dados que permite o rápido acesso a n-gramas de qualquer extensão (*suffix arrays*) ao longo de todo o corpus.

- **Aplicação dos padrões:**
  - **Padrão nominal:** Esse padrão permite extrair uma estrutura contendo NOMINALIZAÇÃO + PREPOSIÇÃO + DETERMINANTE + NOME

    ```
    <pat> <w pos="NOM"/> <w pos="PRP+DET"/> <w pos="NOM" /> </pat>
    ```
  - **Padrão verbal:** Esse padrão permite extrair uma estrutura contendo VERBO + NOME

    ```
    <pat> <w pos="V"/> <w pos="NOM”/> </pat>
    ```

- **Contagem (2)**: cálculo do número de ocorrências do candidato no *corpus* e do número de ocorrências de cada um dos seus componentes individualmente.

- **Medidas de associação (3)** (*Maximum Likelihood Estimator e Student's t test score*): cálculo das ocorrências individuais das palavras que compõem os candidatos e que indica se sua frequência como uma expressão é maior do que a sua frequência como uma palavra isolada.

- **Contraste**: o contraste entre as medidas encontradas no *corpus* de estudo e no *corpus* de referencia.

- **Filtros (4 e 5)**: Aplicação do corte de frequência igual ou superior a dois sobre os resultados obtidos com o contraste, obtendo-se um conjunto mais reduzido de expressões.
  Aplicação de novo filtro linguístico: identificação das expressões a partir dos termos mais frequentes no *corpus* (agrotóxico, água, dano, energia, espécies, produto, recursos e resíduos)

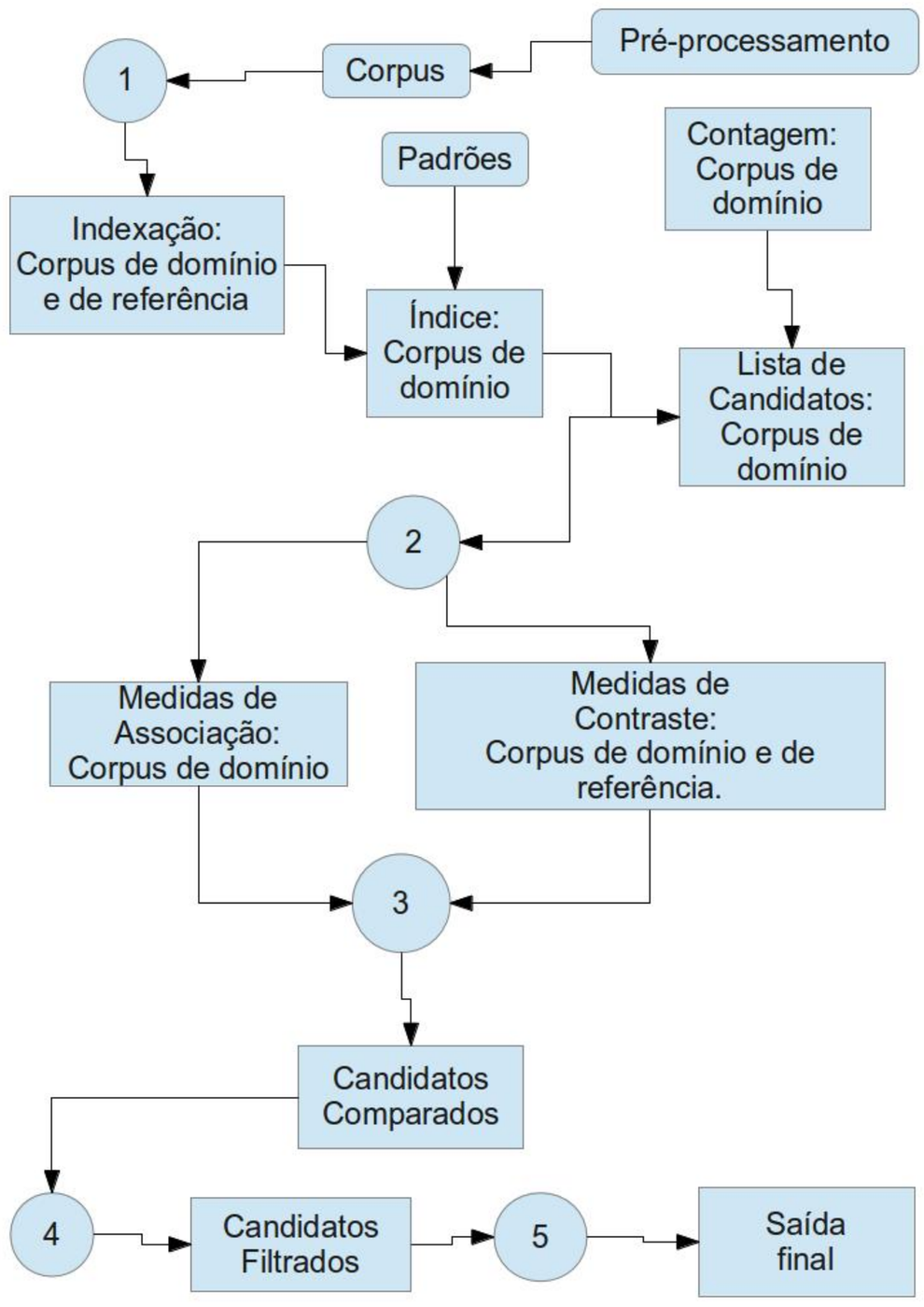

Figura: Etapas da aplicação do `mwetoolkit`

Resultados: O resultado preliminar gerou 7638 candidatos nominais e 3244 candidatos verbais, abrangendo todos os termos. Após a aplicação dos filtros linguísticos e dos critérios estatísticos de contraste o resultado final foi de 225 candidatos nominais e 22 candidatos verbais para o grupo dos oito termos mais comuns no *corpus* na posição `NOM`.

Exemplo de expressões obtidas após o processo:

- acondicionamento do produto (normalização)
- acondicionar resíduos (verbal).

Considerações finais: Após a utilização do `mwetoolkit` para a extração de expressões multivocabulares, pode-se destacar, entre suas principais vantagens, a precisão dos candidatos extraídos, principalmente os formados por padrões verbais.

Referências: RAMISCH, Carlos. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. 2012. Disponível em: http://www.inf.ufrgs.br/~ceramisch/download_files/thesis/. Acesso em 30/08/2012.

FAPERGS
Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
[TERMISUL] TERMINOLÓGICO CONESUL