Aplicação do `mwetoolkit` na pesquisa de Combinatórias Léxicas Especializadas da linguagem legislativa do meio ambiente

Paulo G. P. Duarte, Anna M. B. Maciel, Cleci R. Bevilacqua

Fourth CAMELEON Project Workshop
13/12/2013
Instituto de Informática, UFRGS

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

[TERMINOLÓGICO CONE SUL]

CNPq

Contextualização

- **TERMISUL**

  - Projeto *Combinatórias Léxicas Especializadas (CLEs) da linguagem legal, normativa e científica* (ProjeCOM Legis).

- **Objetivo**

  - Coletar as expressões multivocabulares prototípicas da legislação ambiental brasileira e seus equivalentes na legislação do meio ambiente dos países do Mercosul, e nas leis ambientais da Alemanha, Estados Unidos, França, Itália.

  - Criar uma base de dados *on-line* com as expressões coletadas.

Objetivo desta comunicação

- 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

Materiais

- **Corpus especializado**
  - **LegisAmb_br**
  - 280 diplomas legais de1934 a 2010 da Base Legis do Projeto TERMISUL, UFRGS, baixados como < brLegis.zip>.
  - Extensão total: 513.731 tokens e 20.811 types.
  - http://www.ufrgs.br/termisul/bases_textuais/legis/legislacao_ambiental.php

- **Corpus de língua geral**
  - **PLNBRFULL**
  - 100 milhões de palavras da Folha de São Paulo.
  - www.nilc.icmc.usp.br/plnbr

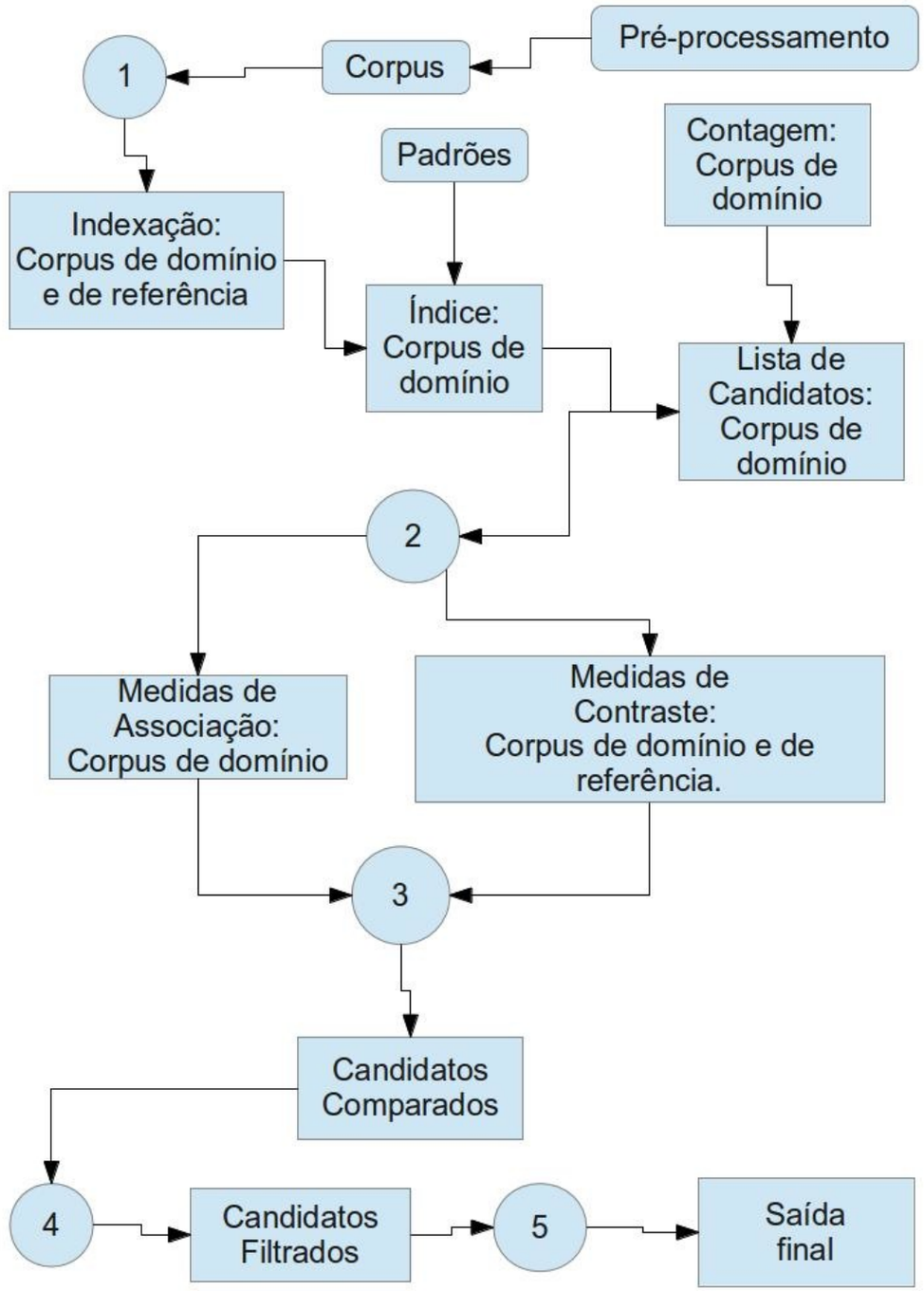
- Lista de padrões de CLEs recorrentes no *corpus* especializado.

mwetoolkit

- Ferramenta desenvolvida por Carlos Ramisch para atuar como extrator de expressões recorrentes formadas por duas ou mais palavras em textos de qualquer língua ou de qualquer linguagem especializada.

- http://mwetoolkit.sourceforge.net/PHITE.php

Aplicação


Pré-processamento
Corpus
1
Padrões
Contagem:
Corpus de
domínio
Indexação:
Corpus de domínio
e de referência
Índice:
Corpus de
domínio
Lista de
Candidatos:
Corpus de
domínio
2
Medidas de
Associação:
Corpus de domínio
Medidas de
Contraste:
Corpus de domínio e de
referência.
3
Candidatos
Comparados
4
Candidatos
Filtrados
5
Saída
final

Pré-processamento
Etiquetagem

- ***Corpus***
  - *<LegisAmb_br.txt>*
- **Ferramenta**
  - TreeTagger (SCHMID, 1994; GAMALLO, 2012)

Tabela 1: Etiquetas do TreeTagger

Adjetivo	ADJ
Advérbio	ADV
Conjunção coordenativa	CONJ
Conjunção Subordinativa	CONJSUB
Determinante	DET
Número Cardinal / Ordinal	CARD
Nome Comum / Próprio	NOM
Particípio passado	PP
Pronome	P
Preposição	PREP
Preposição mais Determinante	PREP+DET
Verbo	V
Verbo mais Preposição	V+P
Separadores dentro da oração	VIRG
Separadores de orações	SENT

Problemas e soluções

- **Problemas**
    - Causados em função da língua (PB), do domínio (Direito Ambiental) e gênero (texto legislativo).
- **Soluções**
    - Normalização *(lowercase);*
    - Conversão dos caracteres acentuados do código latin1 para o código UTF8;
    - Eliminação da marcação catalográfica e macroestrutural dos textos;
    - Eliminação dos símbolos $, %, º, §.

Corpus pré-processado

```
<!DOCTYPE corpus SYSTEM "dtd/mwetoolkit-corpus.dtd">
<corpus >
<s s_id="0">
    <w surface="LEI" pos="NOM" lemma="lei"/>
    <w surface="N" pos="NOM" lemma="n"/>
    <w surface="612" pos="CARD" lemma="@card@"/>
    <w surface="," pos="VIRG" lemma=","/>
    <w surface="DE" pos="NOM" lemma="DE"/>
    <w surface="13" pos="CARD" lemma="@card@"/>
    <w surface="DE" pos="NOM" lemma="DE"/>
    <w surface="JANEIRO" pos="NOM" lemma="janeiro"/>
    <w surface="DE" pos="NOM" lemma="DE"/>
    <w surface="1949" pos="CARD" lemma="@card@"/>
    <w surface="Cria" pos="NOM" lemma="cria"/>
    <w surface="um" pos="DET" lemma="um"/>
    <w surface="Hôrto" pos="NOM" lemma="Hôrto"/>
    <w surface="Florestal" pos="NOM" lemma="Florestal"/>
    <w surface="no" pos="PRP+DET" lemma="em"/>
    <w surface="Município" pos="NOM" lemma="município"/>
    <w surface="de" pos="PRP" lemma="de"/>
    <w surface="Silvânia" pos="NOM" lemma="Silvânia"/>
    <w surface="," pos="VIRG" lemma=","/>
    <w surface="no" pos="PRP+DET" lemma="em"/>
    <w surface="Estado" pos="NOM" lemma="estado"/>
    <w surface="de" pos="PRP" lemma="de"/>
    <w surface="Goiás" pos="NOM" lemma="Goiás"/>
    <w surface="." pos="SENT" lemma="."/>
</s>
```

Lista dos padrões

- Padrões mais recorrentes e menos complexos no corpus especializado.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE dict SYSTEM "dtd/mwetoolkit-patterns.dtd">
<patterns>
    <pat> <w pos="NOM"> <w pos="PRP+DET"> <w pos="NOM"></pat>
    <pat> <w pos="NOM"> <w pos="PRP"> <w pos="NOM"></pat>
    <pat> <w pos="NOM"> <w pos="PRP"> <w pos="NOM"> <w pos="ADJ"> </pat>
</patterns>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE dict SYSTEM "dtd/mwetoolkit-patterns.dtd">
<patterns>
    <pat> <w pos="V"> <w pos="NOM"> </pat>
    <pat> <w pos="V"> <w pos="DET"> <w pos="NOM"> </pat>
    <pat> <w pos="V"> <w pos="PRP+DET"> <w pos="NOM"> </pat>
    <pat> <w pos="V"> <w pos="NOM"> <w pos="ADJ"> </pat>
    <pat> <w pos="V"> <w pos="DET"> <w pos="NOM"> <w pos="PRP+DET"> <w pos="NOM">  </pat>
</patterns>
```

Indexação

- 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 <*taggedLegisAmb_br.xml*> e no PLNBRFULL.

Extração de candidatos a CLEs

patterns.xml ⟶ corpus.xml ⟶ candidates-from-index.xml

Candidatos extraídos

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE candidates SYSTEM "dtd/mwetoolkit-candidates.dtd">
<candidates >
<meta>
</meta>
<cand candid="0">
    <ngram>
        <w surface="cometer" pos="V" />
        <w surface="a" pos="DET" />
        <w surface="funcionários" pos="NOM" />
    </ngram>
    <occurs>
    <ngram>
        <w surface="cometer" lemma="cometer" pos="V" />
        <w surface="a" lemma="a" pos="DET" />
        <w surface="funcionários" lemma="funcionário" pos="NOM" />
        <freq name="corpus" value="1" />
    </ngram>
    </occurs>
</cand>
```

Contagem

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE candidates SYSTEM "dtd/mwetoolkit-candidates.dtd">
<candidates >
<meta>
    <corpussize name="legisamb" value="674696" />
</meta>
<cand candid="0">
        <ngram>
        <w lemma="chefe" pos="NOM" >
                <freq name="legisamb" value="79" />
        </w>
                <w lemma="de" pos="PRP+DET" >
                <freq name="legisamb" value="74322" />
        </w>
                <w lemma="casa" pos="NOM" >
                <freq name="legisamb" value="24" />
        </w>
                <freq name="legisamb" value="7" />
        </ngram>
        <occurs>
        <ngram>
                <w surface="Chefe" lemma="chefe" pos="NOM" />
                <w surface="da" lemma="de" pos="PRP+DET" />
                <w surface="Casa" lemma="casa" pos="NOM" />
                <freq name="legisamb" value="7" />
        </ngram>
    </occurs>
</cand>

```

Medidas associativas

- Medidas associativas:
    - mle: *Maximum Likelihood Estimator*
    - t: *Student's t test score*

Medidas associativas

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE candidates SYSTEM "dtd/mwetoolkit-candidates.dtd">
<candidates >
<meta>
    <corpussize name="corpus" value="674696" />
    <metafeat name="mle_corpus" type="real" />
    <metafeat name="t_corpus" type="real" />
</meta>
<cand candid="0">
    <ngram>
        <w surface="atender" pos="V" >
                <freq name="corpus" value="151" />
        </w>
        <w surface="à" pos="PRP+DET" >
                <freq name="corpus" value="3039" />
        </w>
        <w surface="notificação" pos="NOM" >
                <freq name="corpus" value="111" />
        </w>
                <freq name="corpus" value="1" />
        </ngram>
    <occurs>
        <ngram>
                <w surface="atender" lemma="atender" pos="V" />
                <w surface="à" lemma="a" pos="PRP+DET" />
                <w surface="notificação" lemma="notificação" pos="NOM" />
                <freq name="corpus" value="1" />
        </ngram>
    </occurs>
    <features>
        <feat name="mle_corpus" value="1.48214899747e-06" />
        <feat name="t_corpus" value="0.999888104385" />
    </features>
</cand>
```

Contraste: corpus de domínio e corpus de referência

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE candidates SYSTEM "dtd/mwetoolkit-candidates.dtd">
<candidates >
<meta>
    <corpussize name="corpus" value="674696" />
    <corpussize name="folha.corpus" value="2064512" />
    <metafeat name="csmwe_corpus" type="real" />
    <metafeat name="simplediff_corpus" type="real" />
    <metafeat name="simplecsmwe_corpus" type="real" />
    <metafeat name="mle_corpus" type="real" />
    <metafeat name="t_corpus" type="real" />
    <metafeat name="pmi_corpus" type="real" />
    <metafeat name="dice_corpus" type="real" />
    <metafeat name="ll_corpus" type="real" />
    <metafeat name="mle_folha.corpus" type="real" />
    <metafeat name="t_folha.corpus" type="real" />
    <metafeat name="pmi_folha.corpus" type="real" />
    <metafeat name="dice_folha.corpus" type="real" />
    <metafeat name="ll_folha.corpus" type="real" />
</meta>
<cand candid="0">
    <ngram>
        <w lemma="chefe" pos="NOM" >
                <freq name="corpus" value="79" />
                <freq name="folha.corpus" value="164" />
        </w>
        <w lemma="de" pos="PRP+DET" >
                <freq name="corpus" value="74322" />
                <freq name="folha.corpus" value="139437" />
        </w>
        <w lemma="casa" pos="NOM" >
                <freq name="corpus" value="24" />
                <freq name="folha.corpus" value="875" />
        </w>
                <freq name="corpus" value="7" />
                <freq name="folha.corpus" value="0" />
        </ngram>
    <occurs>
        <ngram>
                <w surface="Chefe" lemma="chefe" pos="NOM" />
                <w surface="da" lemma="de" pos="PRP+DET" />
                <w surface="Casa" lemma="casa" pos="NOM" />
                <freq name="corpus" value="7" />
        </ngram>
    </occurs>

```

Etapas seguintes

- Filtragem dos candidatos para que tenhamos uma quantidade menor de dados para serem analisados. Isso é feito através do próprio mwetoolkit, ou utilizando alguma ferramenta externa, como R ou MS Excel

- Análise dos candidatos filtrados.

Referências

- ANTHONY, L. *AntConc* (*3.2.1 w).* Tokyo: Waseda University, 2008. Disponível em http://www.antlab.sci.waseda.ac.jp/software.html. Acesso em: 28 de abr 2012.

- BEVILACQUA, C.R; MACIEL, A. M. B.; REUILLARD, P. C. R.; SCHEEREN, C. M. Combinatórias Léxicas Especializadas: etapas prévias para identificação e tratamento. XII Simposio Iberoamericano de Terminología RITerm 2010. Actas. Vol. 2. Buenos Aires: Colegio de Traductores de la Ciudad de Buenos Aires, 2012, p. 272-283.

- BEVILACQUA, C. R.; MACIEL, A. M. B.; REUILLARD, P. C. R.; SCHEEREN, C. M.; KILIAN, Cristiane Krause. Combinatórias Léxicas da Linguagem Legislativa: uma abordagem orientada pelo *corpus*. In: MURAKAWA, Clotilde de Almeida Azevedo; SILVA, Odair Luiz (Org.) *Terminologia: uma ciência interdisciplinar.* Araraquara, FCL-UNESP Laboratório Editorial/Cultura Acadêmica Editora, 2013 p. 277-244.

- GAMALLO, Pablo. *PoS Tree-Tagger for Portuguese and Galician*. 2005 Disponível em: < http://gramatica.usc.es/~gamallo/tagger.htm > Acesso em: 02/02/2012.

- KRIEGER, MACIEL, FINATTO, ROCHA, BEVILACQUA. Dicionário de Direito Ambiental: terminologia das leis do meio ambiente Rio de Janeiro: Lexikon, 2008.

- 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.

- RAMISCH, Carlos; VILLAVICENCIO, Aline; BOITET, Christian, 2010 *Framework for Multiword Expression Identification.* Disponível em http://www.lrec-conf.org/proceedings/lrec2010/pdf/803_Paper.pdf Acesso em: 12/03/2012.

- SCHMID, Helmut. Probabilistic Part of Speech Tagging using decision trees, 1994) Disponível em: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf Acesso em: 03/11/2011.