Processamento de Linguagem Natural:

Identificação de Expressões Multipalavra

Vítor De Araújo *vbuaraujo@inf.ufrgs.br*
Prof. Edson Prestes e Silva Jr. *prestes@inf.ufrgs.br*
Profª Aline Villavicencio *avillavicencio@inf.ufrgs.br*
Carlos Ramisch *ceramisch@inf.ufrgs.br*

O que é?

Expressões multipalavra (EMs): combinações de palavras que apresentam idiossincrasias lingüísticas ou estatísticas

- Verbos frasais: *carry up, consist of*
- Verbos de suporte: *tomar um banho*
- Compostos: *carro de polícia, bode expiatório*
- Expressões idiomáticas: *engolir o sapo, dar para trás*

mwetoolkit (*mwetoolkit.sf.net*)**:** ferramenta automatizada para a identificação e extração de EMs a partir de corpora utilizando métodos estatísticos.

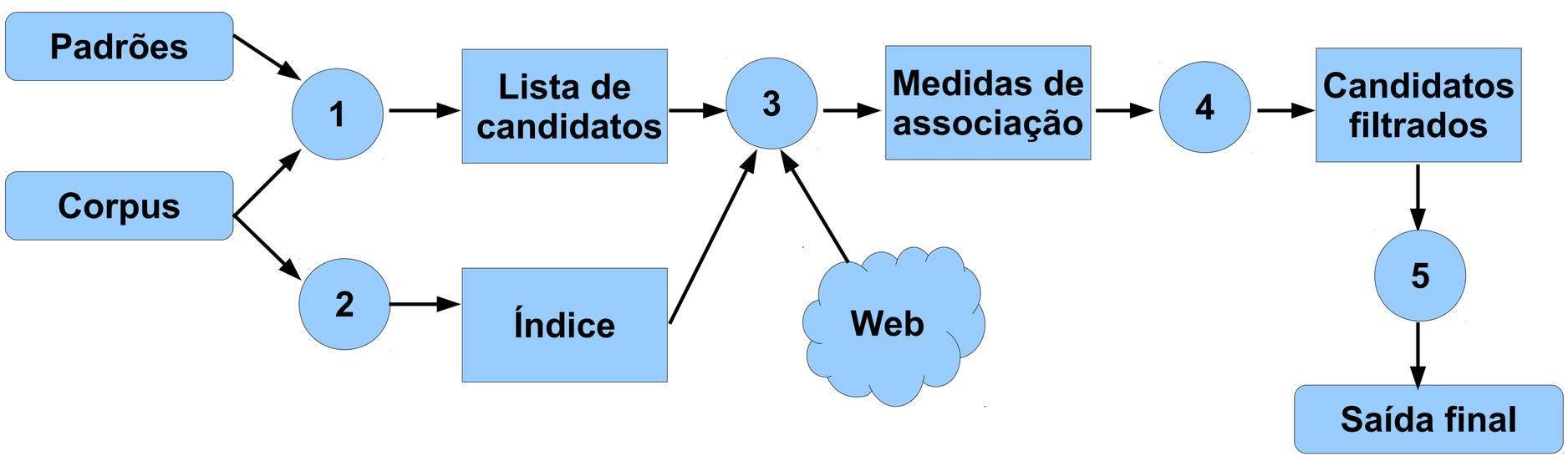
Motivação

Fenômeno **ubíquo** em linguagem natural, devendo ser tratado por ferramentas que realizam processamento semântico (e.g., tradução automática).

Objetivo

Melhorar o desempenho do *mwetoolkit*:
- Eficácia da identificação de EMs
- Consumo de recursos computacionais

Como funciona?



1. Extração: percorrimento do corpus previamente pré-processado em busca de sequências de palavras (n-gramas) que são possíveis EMs de acordo com determinados padrões morfossintáticos (e.g., sequências verbo-substantivo).

2. Indexação: geração de uma *array de sufixos* (estrutura para cálculo eficiente da frequência de n-gramas). Também é possível usar a Web como fonte de frequências, através de mecanismos de busca (e.g., Google).

3. Geração de atributos: cálculo de medidas de associação a partir das frequências de palavras individuais e de n-gramas. Essas medidas estimam a probabilidade de um n-grama ser uma EM.

4. Filtragem: seleção heurística dos candidatos com maior probabilidade.

5. Avaliação: comparação dos candidatos selecionados com um *gold standard*, se disponível. Alternativamente, pode-se usar uma ferramenta de aprendizado de máquina.

Contribuições

- Suporte a **expressões regulares** em padrões
  - Repetições (*substantivo + um ou mais adjetivos*)
  - Itens opcionais (*substantivo, precedido ou não de artigo*)
  - Backreferences (e.g., *dia após dia, passo a passo*)
- Padrões com **dependências sintáticas** (*verbo + objeto*)
- Tratamento de **EMs não-contíguas**
- **Indexação mais eficiente** (menor consumo de tempo e memória)
- **Interface** de comandos unificada para acesso às funcionalidades do *toolkit*
- Uso do Google **Web 1 Trillion 5-Gram** corpus para cálculo de freqüências
- Suporte ao algoritmo **LocalMaxs** para extração de candidatos independente de filtragem
- **Avaliação preliminar** da implementação de expressões regulares e dependências sintáticas sobre estudo de EMs em corpus de transcrição de discurso de crianças (CHILDES)

Trabalhos futuros

- Comparar o *mwetoolkit* com outras ferramentas
- Tratamento de EMs aninhadas
- Melhorar o desempenho da extração de candidatos

Conclusão

Trabalho de melhoria, otimização e avaliação de uma ferramenta de extração de EMs: desafio para PLN

Referências

V. de Araújo, C. Ramisch, A. Villavicencio. *Fast and Flexible MWE Candidate Generation with the mwetoolkit*. In MWE 2011, Portland, Oregon, USA. http://aclweb.org/anthology-new/W/W11/W11-0822.pdf

C. Ramisch, A. Villavicencio, C. Boitet. 2010b. *mwetoolkit: a framework for multi-word expression identification.* In Proc. of the Seventh LREC (LREC 2010), Malta, May. ELRA.
http://inf.ufrgs.br/~ceramisch/download_files/publications/2010/p09.pdfr