

LALIC

Laboratório de Linguística e
Inteligência Computacional

www.lalic.dc.ufscar.br

NEMWEL

Bilingual MWE identification

Helena Caseli

helenacaseli@dc.ufscar.br

October 2016



Núcleo Interinstitucional de
Linguística Computacional



Never-ending Learning at LALIC

- Background

- AIM-WEST project was born in the scope of a FAPESP project which aimed at developing never-ending systems
 1. NEBEL (pt-en) - bilingual dictionary
 2. NEPaL (pt) - monolingual dictionary
 3. NESReL (pt) - monolingual semantic relations (is-a, used-for, property-of, made-of, part-of, location-of, effect-of)
 4. **NEMWEL (pt) - monolingual list of MWE**

Never-ending Learning at LALIC

■ Never-ending learning at LALIC

<http://www.lalic.dc.ufscar.br/never-ending/>

Principal

Pessoas

Publicações

Recursos & Ferramentas

Sobre o projeto

Learners ▾

English ▾

Never-Ending Learning

A quantidade de informação disponível na web, em um ou vários idiomas, motivou o desenvolvimento de aprendizes automáticos implementados neste projeto. Inserido na área de Processamento de Língua Natural (PLN), este projeto foi concebido com o intuito de extrair automaticamente conhecimento útil para aplicações mono e bilíngues. Desse modo, foram desenvolvidos sistemas computacionais capazes de extrair automaticamente e a partir de repositórios e páginas online: **traduções, paráfrases, relações semânticas e expressões multipalavras**. A estratégia investigada foi o **aprendizado de máquina sem-fim** (AMSF ou, do inglês, *never-ending learning*) tendo a web como fonte de conhecimento. O AMSF é uma estratégia de aprendizado de máquina baseada no aprendizado constante e incremental inspirada no modo como nós, humanos, aprendemos.

Visão geral »

Never-ending Learning at LALIC

■ Never-ending learning at LALIC

<http://www.lalic.dc.ufscar.br/never-ending/>

NESReL

👍 🔄 6.333333	capital 🇧🇷	alagoana 🇧🇷	property-of	16/9/2015
<p>Segundo o documento , se em 1991 , apresentavam IDHM " Muito=Baixo " 101 municípios alagoanos e apenas um apresentava índice " Baixo " , em 2010 , os municípios avançaram pontuando apenas 2 de eles com IDHM " Muito=Baixo " , 86 como " Baixo " , 13 como " Médio " , e a <u>capital alagoana</u> , Maceió , com o percentual de IDHM " Alto " .</p>				
👍 🔄 3.333333	Maceió 🇧🇷	capital 🇧🇷	is-a	16/9/2015
<p>Segundo o documento , se em 1991 , apresentavam IDHM " Muito=Baixo " 101 municípios alagoanos e apenas um apresentava índice " Baixo " , em 2010 , os municípios avançaram pontuando apenas 2 de eles com IDHM " Muito=Baixo " , 86 como " Baixo " , 13 como " Médio " , e a <u>capital alagoana</u> , <u>Maceió</u> , com o percentual de IDHM " Alto " .</p>				

Never-ending Learning at LALIC

- Never-ending learning at LALIC

<http://www.lalic.dc.ufscar.br/never-ending/>

NEPaL

- 👍 🔄 42.857 [ave](#) 🇧🇷 [pássaro](#) 🇧🇷 14/9/2015
troçadas por tucanos a aquisição dos animais foi uma permuta . o aquário de dallas doou as girafas e , em troca , recebeu 32 [aves](#) , entre elas tucanos e araçaris .
- 👍 🔄 28.729 [habitar](#) 🇧🇷 [viver](#) 🇧🇷 14/9/2015
para o geólogo luiz carlos borges ribeiro , do centro de pesquisas paleontológicas llewellyn ivor price , ligado a uftm (universidade federal do triângulo mineiro) , a estrutura pode pertencer a uma nova espécie de titanossauro , um dinossauro herbívoro que [habitou](#) a região .

Never-ending Learning at LALIC

■ Never-ending learning at LALIC

<http://www.lalic.dc.ufscar.br/never-ending/>

NEMWEL

- | | | | | |
|--|---|-----------|---|-----------|
| 👍 | 👎 | 26.561367 | partido de oposição 🇧🇷 | 15/9/2015 |
| Os presidentes da Câmara e do Senado disseram que não têm motivos para devolver a proposta para o governo , apesar do pedido de partido de oposição . | | | | |
| 👍 | 👎 | 21.429613 | pauta de reivindicação 🇧🇷 | 15/9/2015 |
| Pauta de reivindicação Também estão entre as solicitações da categoria , segundo o sindicato , o pagamento do percentual de 5 % , referente a última parcela do Plano de Cargos e Salários da Categoria , pagamento da gratificação de atividade externa dos oficiais de Justiça e pagamentos das substituições e acúmulos de cargos . | | | | |
| 👍 | 👎 | 27.391026 | crise econômico 🇧🇷 | 15/9/2015 |
| O presidente , Renan Calheiros , usou uma inspiração pop para pregar o apoio de todos os partidos para superar a crise econômica . | | | | |

NEMWEL

- How it works
 - Source
 - pt: G1 news portal
 - <http://g1.globo.com>
 - pt-en: Folha de São Paulo
 - <http://www1.folha.uol.com.br/internacional>

The screenshot displays the Folha de S. Paulo website interface. At the top, there are navigation options for 'NOTÍCIAS EN ESPAÑOL' and 'NOTÍCIAS EM PORTUGUÊS', along with social media follow buttons for Twitter and Facebook. The main headline reads 'Brazil Sends Students Abroad Too President of Scientific Agency', dated 08/15/2016 - 11H44. The author is identified as SABINE RIGHETTI, a contributor from Brasília. The article text states: 'Appointed during the Michel Temer administration, sociologist Abílio Neves, the new president of Capes, Brazil's main agency for scientific advancement, is feeling at home.' To the right, there is a video player showing Abílio Neves, with a caption: 'Abílio Neves, presidente da Capes, principal agência que financia ciência no país'. Below the video, a social media post from Sabine Righetti is visible, dated 14/08/2016 at 02h02, with 2.4 million views. The post includes sharing options for Facebook, Twitter, and LinkedIn, and a link to 'OUVIR O TEXTO'. At the bottom, there are additional sharing options like 'Enviar por e-mail', 'Copiar url curta', 'Imprimir', and 'Comunicar erros'.

NEMWEL

- How it works

- Step-by-step

1. Crawler

- 40 news texts collected in each iteration

2. Extractor

- TreeTagger annotates the lemma and PoS for each token

- MWEtoolkit extracts the candidates

3. Processor

- MWEtoolkit calculates some association measures

- Other features: translatability and PoS context

4. Promoter

- SVM model classifies a candidate as true MWE

NEMWEL

- Examples of MWE extracted
 - Portuguese
 - cota racial
 - margem de erro
 - deficit fiscal
 - English
 - racial quota
 - margin of error
 - fiscal deficit

NOW:
extracted separately

NEMWEL

- TO DO

- Find the parallelism between MWE candidates
 - Based on word-alignment (e.g. Giza++)
 - Based on bilingual similarity (e.g. Multivec)

Portuguese

English

cota racial

<-> racial quota

margem de erro

<-> margin of error

deficit fiscal

<-> fiscal deficit

LALIC

Laboratório de Linguística e
Inteligência Computacional

www.lalic.dc.ufscar.br

NEMWEL

Bilingual MWE identification

Helena Caseli

helenacaseli@dc.ufscar.br

October 2016



Núcleo Interinstitucional de
Linguística Computacional

