NEMWEL
Bilingual MWE identification

Helena Caseli

helenacaseli@dc.ufscar.br
October 2016

Núcleo Interinstitucional de Linguística Computacional
FAPESP
UNIVERSIDADE FEDERAL DE SÃO CARLOS

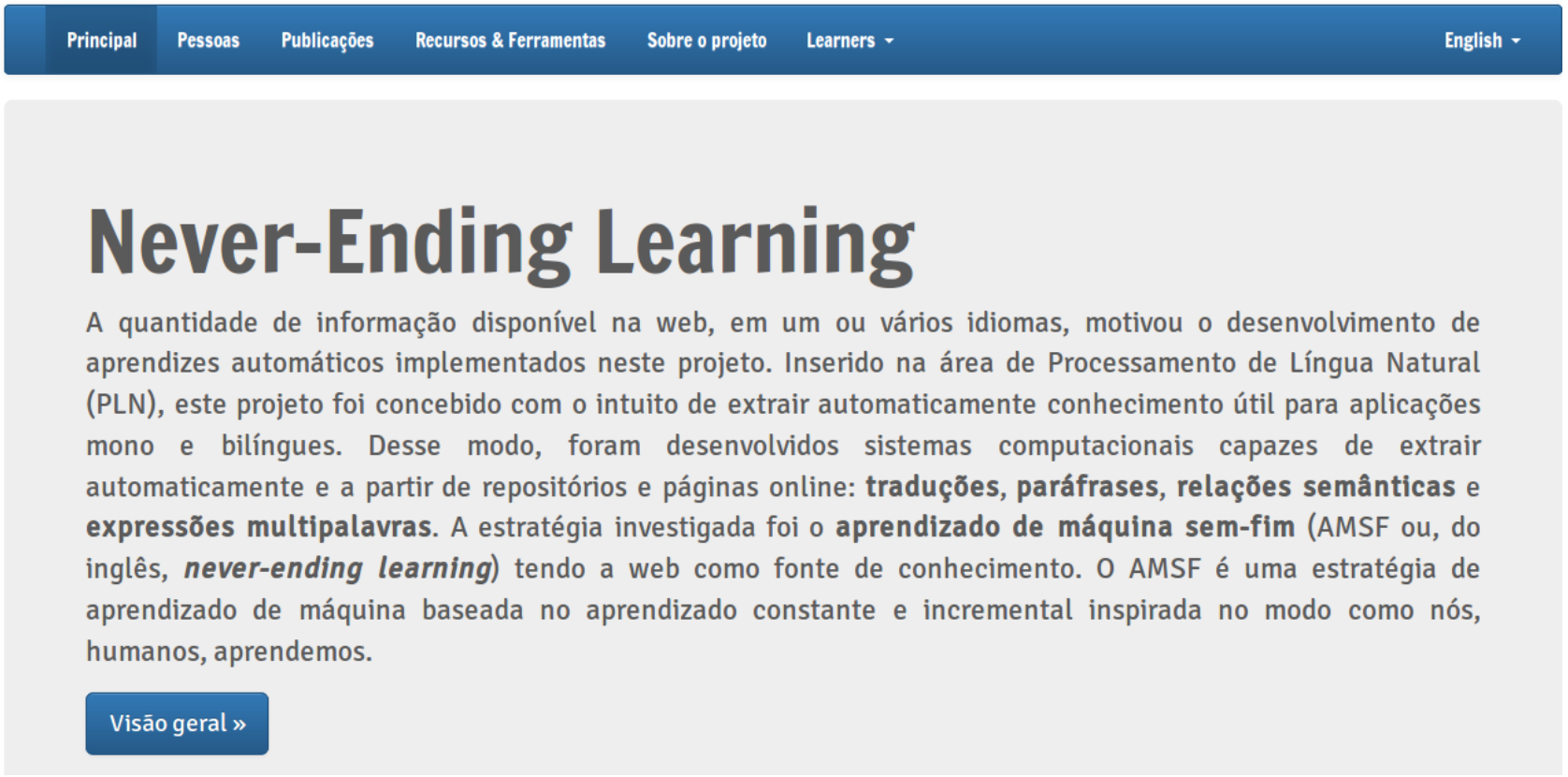
Never-ending Learning at LALIC

- 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/



Never-ending Learning at LALIC

- Never-ending learning at LALIC

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

NESReL

6.333333	capital 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 capital alagoana , Maceió , com o percentual de IDHM " Alto " .	alagoana	property-of	16/9/2015
3.3333333	Maceió 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 capital alagoana , Maceió , com o percentual de IDHM " Alto " .	capital	is-a	16/9/2015

Never-ending Learning at LALIC

- Never-ending learning at LALIC

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

NEPaL

42.857	ave trocadas 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 .	pássaro	14/9/2015
28.729	habitar 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 .	viver	14/9/2015

Never-ending Learning at LALIC

- Never-ending learning at LALIC

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

NEMWEL

26.561367	partido de oposição 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 .	15/9/2015
21.429613	pauta de reivindicação 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 .	15/9/2015
27.391026	crise econômico O presidente , Renan Calheiros , usou uma inspiração pop para pregar o apoio de todos os partidos para superar a crise econômica .	15/9/2015

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

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

NEMWEL
Bilingual MWE identification

Helena Caseli

helenacaseli@dc.ufscar.br
October 2016

Núcleo Interinstitucional de Linguística Computacional
FAPESP
UNIVERSIDADE FEDERAL DE SÃO CARLOS