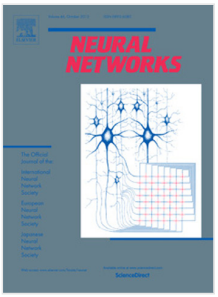
Accepted Manuscript



A universal multilingual weightless neural network tagger via quantitative linguistics

Hugo C.C. Carneiro, Carlos E. Pedreira, Felipe M.G. França, Priscila M.V. Lima

PII: S0893-6080(17)30091-6
DOI: http://dx.doi.org/10.1016/j.neunet.2017.04.011
Reference: NN 3750

To appear in: *Neural Networks*

Received date: 25 November 2016
Revised date: 4 March 2017
Accepted date: 18 April 2017

Please cite this article as: Carneiro, H. C. C., Pedreira, C. E., França, F. M. G., & Lima, P. M. V. A universal multilingual weightless neural network tagger via quantitative linguistics. *Neural Networks* (2017), http://dx.doi.org/10.1016/j.neunet.2017.04.011

A universal multilingual weightless neural network tagger via quantitative linguistics

Hugo C. C. Carneiro[a], Carlos E. Pedreira[a], Felipe M. G. França[a], Priscila M. V. Lima[b]

[a] *Systems Engineering and Computer Science Program/COPPE, Universidade Federal do Rio de Janeiro (UFRJ) - Caixa Postal 68511, Cidade Universitária, Rio de Janeiro, Rio de Janeiro 21941-972, Brazil*
[b] *Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais (NCE), Universidade Federal do Rio de Janeiro (UFRJ) - Av. Athos da Silveira Ramos, 274 - Edifício do Centro de Ciências Matemáticas e da Natureza, Bloco E, Cidade Universitária, Rio de Janeiro, Rio de Janeiro 21941-916, Brazil*

Abstract

In the last decade, given the availability of corpora in several distinct languages, research on multilingual part-of-speech tagging started to grow. Amongst the novelties there is mWANN-Tagger (**m**ultilingual **w**eightless **a**rtificial **n**eural **n**etwork **tagger**), a weightless neural part-of-speech tagger capable of being used for mostly-suffix-oriented languages. The tagger was subjected to corpora in eight languages of quite distinct natures and had a remarkable accuracy with very low sample deviation in every one of them, indicating the robustness of weightless neural systems for part-of-speech tagging tasks. However, mWANN-Tagger needed to be tuned for every new corpus, since each one required a different parameter configuration. For mWANN-Tagger to be truly multilingual, it should be usable for any new language with no need of parameter tuning. This article proposes a study that aims to find a relation between the lexical diversity of a language and the parameter configuration that would produce the best performing mWANN-Tagger instance. Preliminary analyses suggested that a single parameter configuration may be applied to the eight aformentioned languages. The mWANN-Tagger instance produced by this configuration was as accurate as the language-dependent ones obtained through tuning. Afterwards, the weightless neural tagger was further subjected to new corpora in languages that range from very isolating to polysynthetic ones. The best performing instances of mWANN-Tagger are again the ones produced by the universal parameter configuration. Hence, mWANN-Tagger can be applied to new corpora with no need of parameter tuning, making it a universal multilingual part-of-speech tagger. Further experiments with Universal Dependencies treebanks reveal that mWANN-Tagger may be extended and that it has potential to outperform most state-of-the-art part-of-speech taggers if better word representations are provided.

Keywords:
Part-of-speech tagging, Weightless neural networks, Zipf's law, Lexical diversity