Part-of-Speech Tagging of Twitter Microposts only using Distributed Word Representations and a Neural Network

Fréderic Godin, Wesley De Neve, and Rik Van de Walle

In: CLIN 2015, book of abstracts for the 25th meeting of computational linguistics in the Netherlands, 45, 2015.

http://www.clips.uantwerpen.be/~ben/sites/default/files/book_of_abstracts_final.pdf

Part-of-Speech Tagging of Twitter Microposts only using Distributed Word Representations and a Neural Network

Fréderic Godin
Multimedia Lab
Ghent University - iMinds
Ghent, Belgium
frederic.godin@ugent.be

Wesley De Neve
Multimedia Lab & IVY Lab
Ghent University & KAIST
Ghent, Belgium & Daejeon, South Korea
wesley.deneve@ugent.be

Rik Van de Walle
Multimedia Lab
Ghent University - iMinds
Ghent, Belgium
rik.vandewalle@ugent.be

Abstract

Many algorithms for natural language processing rely on manual feature engineering. However, manually finding effective features is a labor-intensive task. Moreover, whenever these algorithms are applied on new types of content, they do not perform that well anymore and new features need to be engineered. For example, current algorithms developed for Part-of-Speech (PoS) tagging of news articles with Penn Treebank tags perform poorly on microposts posted on social media. As an example, the state-of-the-art Stanford tagger trained on news article data reaches an accuracy of 73% when PoS tagging microposts. When the Stanford tagger is retrained on micropost data and new micropost-specific features are added, an accuracy of 88.7% can be obtained.

We show that we can achieve state-of-the-art performance for PoS tagging of Twitter microposts by solely relying on automatically inferred distributed word representations as features and a neural network. To automatically infer the distributed word representations, we make use of 400 million Twitter microposts. Next, we feed a context window of distributed word representations around the word we want to tag to a neural network to predict the corresponding PoS tag. To initialize the weights of the neural network, we pre-train it with large amounts of automatically high-confidence labeled Twitter microposts. Using a data-driven approach, we finally achieve a state-of-the-art accuracy of 88.9% when tagging Twitter microposts with Penn Treebank tags.