Chapter 1

Embeddings.jl: easy access to pretrained word embeddings from Julia

This paper is currently under review for the Journal of Open Source Software.

1.1 Summary

Embeddings.jl is a tool to help users of the Julia programming language (**Julia**) make use of pretrained word embeddings for NLP. Word embeddings are a very important feature representation in natural language processing. The use of embeddings pretrained on very large corpora can be seen as a form of transfer learning. It allows knowledge of lexical semantics derived from the distributional hypothesis– that words occurring in similar contexts have similar meaning– to be injected into models which may have only limited amounts of supervised, task oriented training data.

Many creators of word embedding methods have generously made sets of pretrained word representations publicly available. Embeddings.jl exposes these as a standard matrix of numbers and a corresponding array of strings. This lets Julia programs use word embeddings easily, either on their own or alongside machine learning packages such as Flux (**flux**). In such deep learning packages, it is common to use word embeddings as an input layer of an LSTM or other neural network, where they may be kept invariant or used as initialization for fine-tuning on the supervised task. They can be summed to represent a bag of words, concatenated to form a matrix representation of a sentence or document, or used otherwise in a wide variety of natural language processing tasks.

Embeddings.jl makes use of DataDeps.jl (**2018arXiv180801091W**), to allow for convenient automatic downloading of the data when and if required. It also uses the DataDeps.jl prompt to ensure the user of the embeddings has full knowledge of the original source of the data, and which papers to cite etc.

It currently provides access to

- multiple sets of word2vec embeddings (**mikolov2013efficient**) for English
- multiple sets of GLoVE (**pennington2014glove**) embeddings for English
- multiple sets of FastText embeddings (**bojanowski2016enriching**; **fasttext157lang**) for several hundred languages

It is anticipated that as more pretrained embeddings are made available for more languages and using newer methods, the Embeddings.jl package will be updated to

support them.