Polyglot2 Documentation

Release 2.0

Rami Al Rafou'

November 21, 2016

Contents

Polyglot2 is a package which allows one to build language models that learn distributed representations of words.

Distributed representations of words (very popularly known as *word embeddings*) have been shown to be useful as features for several Natural Language Processing Tasks like Named Entity Recognition etc.

Polyglot2 allows you to create your own embeddings from text. It's a piece of cake really. Try it out !.

CHAPTER 1

What is Polyglot2 ?

With deep learning taking off with a bang, learning representations from unsupervised data has been an exciting area of research with several applications including the field of Computer Vision, Natural Language Processing etc. In their seminal work Natural Language Processing (almost) from scratch Ronnan Colbert, Jason Weston and others demonstrated that using distributed word representations could achieve competitive and even state of the art results on several natural language processing tasks like part of speech tagging etc. They outline their system SENNA here.

Polyglot2 implements a language model that learns *word embeddings* using a very similar approach as outlined on the above paper. We in fact provide embeddings for more than 100 languages. We encourage you to take a look at them at http://bit.ly/embeddings.

If you would like to train your own embeddings on a corpus, Polyglot2 allows you do that very easily.

CHAPTER 2

Installation guide

2.1 Pre-requisites

The installation steps assume that you have the following things installed:

2.1.1 Python 2.7

Python 3.0 support is coming soon.

2.1.2 BLAS

Our library requires cblas headers to compile. Ubuntu comes with two options:

- OpenBLAS:

```
sudo apt-get install libopenblas-base  libopenblas-dev
```

- ATLAS:

```
sudo apt-get install libatlas3gf-base libatlas-dev
```

The choice of BLAS library influences greatly the speed of the training. We recommend using OpenBLAS as it proved to be faster than ATLAS by **4x**. In case, you want to compile your own OpenBLAS or switch between the two libraries, we compiled a list of useful commands for that purpose in this Tutorial.

2.1.3 cython

Cython is optional but recommended.

2.2 Installing Polyglot2

Download the source from https://bitbucket.org/aboSamoor/polyglot2

Install by using the following command:

```
python setup.py install
```

You are all set !

CHAPTER 3

First Steps

3.1 Pre-requisites

We assume you have successfully installed polyglot2 (Yay!). Here are some very quick steps to get started

3.2 Training a model

To get you started on training a model, we have already provided a sample script that reads in a text corpus and trains a model. The script **polyglot2_trainer.py** in **scripts** allows you to train a simple model as below::

```
polyglot2_trainer.py --files <text files> --output <model file>
```

You can just run the below to get a list of command line options:

```
polyglot2_trainer.py --help
```

3.3 Examining the embeddings

Its very easy to load the trained model and examine the embeddings that have been learnt. If you are familiar with the *word2vec* module in gensim the interface is very similar.

To illustrate assume we have a model named **test.model** trained on some text. We can easily load the model as below:

```
In [1]: from polyglot2 import Polyglot
In [2]: model = Polyglot.load_word2vec_format('test.model')
```

After loading the model, one can easily query the nearest words to a given word(based on thier euclidean distance in embedding space).:

```
In [3]: model.most_similar('king')
Out[4]:
        [(u'king', 0.0),
         (u'prince', 0.58161491620762695),
         (u'queen', 0.61713733694058359),
         (u'emperor', 0.61844666306850182),
         (u'lord', 0.64116868440576313),
         (u'president', 0.66686299825558359),
         (u'captain', 0.702852998721334),
         (u'prophet', 0.72744270206467843),
```

```
        (u'pope', 0.73201129536853193),
        (u'governor', 0.74257922097558959)]
```

As you can see we get the words most similar to *king* in increasing order of their distances.

For more details please take a look at the source here: https://bitbucket.org/aboSamoor/polyglot2

Have fun training and exploring your own word embeddings !

CHAPTER 4

Citing Polyglot

If you use Polyglot, it would be great if you cite us as follows via BIBTEX::

```
@InProceedings{polyglot:2013:ACL-CoNLL,
 author    = {Al-Rfou, Rami  and  Perozzi, Bryan  and  Skiena, Steven},
 title     = {Polyglot: Distributed Word Representations for Multilingual NLP},
 booktitle = {Proceedings of the Seventeenth Conference on Computational Natural Language Learning},
 month     = {August},
 year      = {2013},
 address   = {Sofia, Bulgaria},
 publisher = {Association for Computational Linguistics},
 pages     = {183--192},
 url       = {http://www.aclweb.org/anthology/W13-3520}
}
```

CHAPTER 5

Contact us

We would love to hear from about how Polyglot2 has been useful to you. If you face issues, or have requests please contact us.