

AI Language Technology Lab

Valeriia Volkovaia

April 2019

1 Training Russian nballs

1.1 Requirements

For training nballs for Russian language the following resources were used:

1. Pre-trained word vectors (word2vec) for Russian language: <https://github.com/Kyubyong/wordvectors>
Please extract the file in zip format and save file ru.tsv
2. Russian wordnet <https://wiki-ru-wordnet.readthedocs.io/en/latest/> in the form of package for Python. The documentation is provided in Russian by the link above.
The package can be installed by pip with a command:
pip install wiki-ru-wordnet

1.2 Step 1: convert word2vec file from tsv to standard format

Please run file format_w2v_file.py. The version of python is 3.5. The file ru.tsv has to be placed in the same directory as the .py file. The code produces an output file ru_w2v.txt which contains word2vec features in the form:

```
word_1 feature_1 feature_2 ... feature_256
word_2 feature_1 feature_2 ... feature_256
.
.
.
word_50102 feature_1 feature_2 ... feature_256
```

1.3 Step 2: Create catcode and word sense children files

Please run file make_russian_dataset.py. The version of python is 3.5. The file ru.tsv has to be placed in the same directory as the .py file. The code produces 3 files:

1. idx.dat. The file contains index of word sense and it's definition (in Russian).

2. children.dat_no_duplicates. The file contains the word and it's children. All words without parents are children of *root*.
3. catcode.dat_no_duplicates. The file contains the word and sequence of indexes of it's parents starting from *root*.

1.4 Step 3: Train nballs

Train Russian nballs by the procedure <https://github.com/gnodisnait/nball4tree>. The training command is

```
$ python nball.py --train_nball nball.txt --w2v ru_w2v.txt --ws_child
  children.dat_no_duplicates --ws_catcode catcode.dat_no_duplicates
  --log log.txt
```

Before starting the training please change initialization of first child in the following file:

main_training_process.py:

Line 418:

child0= 'время.н.3'

Line 517:

```
def make_DC_for_first_level_children ( root="*root*", firstChild = 'впе-
  мя.н.3', wsChildrenDic=dict(), outputPath="", maxsize=0, mindim=0, word2ballDic
  = dict(),logFile=None)
```

Line 564:

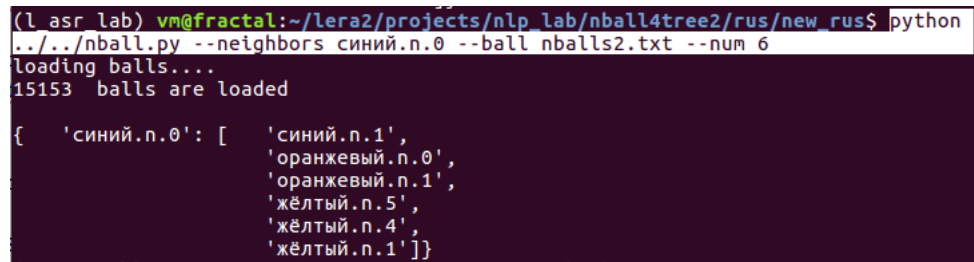
```
make_DC_for_first_level_children(root=root, firstChild = 'время.н.3', wsChildrenDic=wsChildrenDic,
  word2ballDic=word2ballDic, outputPath=outputPath, maxsize=maxsize, mindim=mindim,
  logFile=logFile)
```

2 Results

nballs were tested by finding the neighbors. Here are the some examples:

```
python nball.py --neighbors синий.н.0 --ball nballs2.txt --num 6
```

Output: Tested word sense синий.н.0 meaning blue (color) and closest neighbors



```
(l asr lab) vm@fractal:~/lera2/projects/nlp_lab/nball4tree2/rus/new_rus$ python
../../nball.py --neighbors синий.н.0 --ball nballs2.txt --num 6
loading balls....
15153 balls are loaded
{  'синий.н.0': [  'синий.н.1',
                   'оранжевый.н.0',
                   'оранжевый.н.1',
                   'жёлтый.н.5',
                   'жёлтый.н.4',
                   'жёлтый.н.1']}]
```

are синий.н.1 (blue.н.1), оранжевый.н.0 (orange.н.0), оранжевый.н.1 (orange.н.1),

жёлтый.п.5 (yellow.п.5), жёлтый.п.4 (yellow.п.4), жёлтый.п.1 (yellow.п.1). As we can see, all neighbors are colors as well.

In second example the tested word is март.п.0 (March as a month).
python ../../nball.py --neighbors март.п.0 --ball nballs2.txt --num 6

```
(l_asr_lab) vm@fractal:~/lera2/projects/nlp_lab/nball4tree2/rus/new_rus$ python
../../nball.py --neighbors март.п.0 --ball nballs2.txt --num 6
loading balls....
15153 balls are loaded

{  'март.п.0': [  'ноябрь.п.0',
                  'апрель.п.0',
                  'июнь.п.0',
                  'февраль.п.0',
                  'сентябрь.п.0',
                  'июль.п.0']}]
```

The closest neighbors are: ноябрь.п.0 (November), апрель.п.0 (April), июнь.п.0 (June), февраль.п.0 (February), сентябрь.п.0 (September), июль.п.0 (July), i.e. other monthes.

Next test word is кофе.п.0 (coffee). python ../../nball.py --neighbors кофе.п.0 --ball nballs2.txt --num 6 The closest neighbors are: кофе.п.2 (coffee.п.2),

```
(l_asr_lab) vm@fractal:~/lera2/projects/nlp_lab/nball4tree2/rus/new_rus$ python
../../nball.py --neighbors кофе.п.0 --ball nballs2.txt --num 6
loading balls....
15153 balls are loaded

{  'кофе.п.0': [  'кофе.п.2',
                  'кофе.п.1',
                  'виски.п.0',
                  'чай.п.2',
                  'чай.п.6',
                  'чай.п.0']}]
```

кофе.п.1 (coffee.п.1), виски.п.0 (whiskey.п.0), чай.п.2 (tea.п.2), чай.п.6 (tea.п.6), чай.п.6 (tea.п.6) which are drinks as well.

Another test word is футбол.п.0 (football.п.0):
python ../../nball.py --neighbors футбол.п.0 --ball nballs2.txt --num 6

```
(l_asr_lab) vm@fractal:~/lera2/projects/nlp_lab/nball4tree2/rus/new_rus$ python
../../nball.py --neighbors футбол.п.0 --ball nballs2.txt --num 6
loading balls....
15153 balls are loaded

{  'футбол.п.0': [  'теннис.п.0',
                   'баскетбол.п.0',
                   'бокс.п.4',
                   'бокс.п.2',
                   'бокс.п.1',
                   'бокс.п.3']}]
```

The closest neighbors are: теннис.п.0 (tennis.п.0), баскетбол.п.0 (basketball), бокс.п.2 (boxing.п.2), бокс.п.4 (boxing.п.4), бокс.п.3 (boxing.п.3) which are sports as well.