Recurrent Neural Networks using TensorFlow

Jindřich Libovický

December 5, 2018

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

The Taks

Train a model that will correctly decide whether you should write ‘i’ or ‘y’ in a Czech sentence.

Prepare Python Environment

- create a new environment

  ```
  virtualenv -p python3 env
  ```

- activate the environment

  ```
  source env/bin/activate
  ```

- install TensorFlow and Jupyter

  ```
  pip3 install tensorflow
  pip3 install jupyter
  ```

Alternatively using Anaconda

- download and install Anaconda

```
wget http://repo.continuum.io/archive/Anaconda3-4.2.0-Linux-x86_64.sh
bash Anaconda3-4.2.0-Linux-x86_64.sh
export PATH=$PATH:$HOME/anaconda3/bin
```

- create a new environment

```
conda create -n tf python=3.5 anaconda
```

- activate the environment

```
source activate tf
```

- install TensorFlow 0.12

```
pip3 install tensorflow
```

```
pip3 install jupyter
```

Download the Lab Notebook

- download the notebook

  ```
  wget http://ufallab.ms.mff.cuni.cz/~libovicky/ctu_lab.ipynb
  ```

- run jupyter in the same directory

  ```
  jupyter notebook
  ```

The Data

- 500k sentences from Czech Wikipedia (in general the more, the better)
- only character from Czech alphabet, sentence-splitted, lower-cased
- randomly shuffled, seperated validation data

The text:

```
aristotelés dále určil poloměr země, který ale odhadl na dvojnásobek…
v aristotelovském modelu země stojí a měsíc se sluncem a hvězdami krouží…
mišlenki aristotelovi rozvinul ve 2. století našeho letopočtu klaudios…
```

Correct solution:

```
0000100000000000000000000000000001000001000000000000000000000001000000100000…
0200000200010000000020000010000000000000000010000000000000000000000001000000000…
0000000000000000001000000000001000002000000000000000000020000000000000000000…
```

$1 =$ ‘i’, $2 =$ ‘y’, $0 =$ ‘others’

Baselines

- just leave ‘i’ everywhere
  70 %
- simple rules: ‘y’ after ‘h’, ‘k’, ‘r’ and for words starting with ‘v’
  80 %
- remember the most frequent spelling for each word
  90 %

Learning curves