Unsupervised Learning of Multiple Languages Using Recurrent Neural Networks

Miquel Perelló Nieto, Mathias Berglund\(^1\) and Tapani Raiko\(^1\)

Course:
T-61.5910 Research Project in Computer and Information Science

Aalto, Nov 2013
Index

1. Introduction
   - Motivation

2. Method
   - Summary
   - Corpus
   - Techniques

3. Experiment
   - Experiment

4. Results
   - Error measure
   - Training Error
   - Test Error

5. Discussion
Index

1 Introduction
   • Motivation

2 Method
   • Summary
   • Corpus
   • Techniques

3 Experiment
   • Experiment

4 Results
   • Error measure
   • Training Error
   • Test Error

5 Discussion
Learning multiple languages

- Le langage est la capacité d'exprimer une pensée et de communiquer au moyen d'un système de signes.
- Un idioma ye una llingua, o seya, un sistema de comunicación verbal propio d'una comunidad humana, usáu por ún o varios pueblos o naciones.
- El llenguatge es la facultat de poder comunicar els propis pensaments o sentiments a un receptor o interlocutor mitjançant un sistema o codi determinat de signes interpretable per a ell.
Text prediction

- Involves improving text compression
- Good compression requires a deep understanding of the text
- It can help on human-computer interaction
Deep Neural Networks

- Outstanding in recent challenges
- Ability to get underlying information
- New approaches to train DNN and RNN

---

1 Image from Honglak Lee slides: Deep Learning Methods for Vision

Miquel Perelló Nieto (Aalto)  Writing with RNN  Aalto, Nov 2013
Recent results

- Learned *linguistic and grammatical* structure
- *Balance* parentheses and quotes (e.g., 30 characters)
- Creates *plausible words*
- *Easy to improve* adding more neurons

Example (trained with Wikipedia) \(^2\):

**In**: The meaning of life is

**Out**: *the tradition of the ancient human reproduction: it is less favorable to the good boy[...]*

---

\(^2\)Generating Text with Recurrent Neural Networks[1]
Index

1. Introduction
   - Motivation

2. Method
   - Summary
   - Corpus
   - Techniques

3. Experiment
   - Experiment

4. Results
   - Error measure
   - Training Error
   - Test Error

5. Discussion
Summary

- Create or get a Corpus
- Create N-grams from the Corpus
- Generate and evaluate text with N-grams
- Generate text with RNN
- Compare both systems
Writing with RNN

Method

Summary

Timeline

<table>
<thead>
<tr>
<th>Look for a project</th>
<th>3d</th>
</tr>
</thead>
<tbody>
<tr>
<td>General information</td>
<td>7d</td>
</tr>
<tr>
<td>Overview Neural Networks</td>
<td>3d</td>
</tr>
<tr>
<td>Backpropagation</td>
<td>2d</td>
</tr>
<tr>
<td>Python library: Theano</td>
<td>2d</td>
</tr>
<tr>
<td>Topic presentation</td>
<td></td>
</tr>
<tr>
<td>Data</td>
<td>8d 2h</td>
</tr>
<tr>
<td>Collect Corpus</td>
<td>5d</td>
</tr>
<tr>
<td>Clean Corpus</td>
<td>3d 2h</td>
</tr>
<tr>
<td>N-grams system</td>
<td>15d</td>
</tr>
<tr>
<td>Extract N-grams</td>
<td>2d 3h</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>2d 6h</td>
</tr>
<tr>
<td>Add-One Smoothing</td>
<td>1d 1h</td>
</tr>
<tr>
<td>Add-a Smoothing</td>
<td>1d</td>
</tr>
<tr>
<td>Back-Off</td>
<td>2d</td>
</tr>
<tr>
<td>Interpolation</td>
<td>2d</td>
</tr>
<tr>
<td>Create an n-gram predictor system</td>
<td>3d 5h</td>
</tr>
<tr>
<td>Progress Report</td>
<td></td>
</tr>
<tr>
<td>Recurrent Neural Network</td>
<td>22d 4h</td>
</tr>
<tr>
<td>Create initial NN with Theano</td>
<td>6d 2h</td>
</tr>
<tr>
<td>Create input structure from text</td>
<td>3d 2h</td>
</tr>
<tr>
<td>Training</td>
<td>5d 3h</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>2d</td>
</tr>
<tr>
<td>Mini-batch</td>
<td>3d 3h</td>
</tr>
<tr>
<td>Validate results</td>
<td>7d 4h</td>
</tr>
<tr>
<td>Progress Report</td>
<td></td>
</tr>
<tr>
<td>Write final report</td>
<td>5d</td>
</tr>
<tr>
<td>Final Seminar A</td>
<td></td>
</tr>
</tbody>
</table>
Index

1 Introduction
   - Motivation

2 Method
   - Summary
   - Corpus
   - Techniques

3 Experiment
   - Experiment

4 Results
   - Error measure
   - Training Error
   - Test Error

5 Discussion
By language

**English - 1.4GB**
- Wikipedia
- Previously cleaned

**Spanish - 466MB**
- Joint Research Center
- “Total body of European Union (EU) law applicable in the EU Member States”
- Divided by years in xml format (1958-2006)
- Merged all contents into one file
- Removed accents, “ñ” and “ü”
Writing with RNN

Char frequencies
• Only *kept words* of less than 40 characters
• Larger ones are usually URL’s or numbers
Length sentences

- Removed sentences of less than 50 characters
- also larger than 1000
Writing with RNN
Method

Techniques

N-grams

• Need to choose the $N$
• Preprocess to create the list of N-grams
• Compute frequencies and create a DB
• Smoothing techniques to improve likelihood
  ▶ Add-one Smoothing
  ▶ Add-$\alpha$ Smoothing
  ▶ Good-Turing Smoothing
  ▶ Interpolation
Recurrent Neural Networks

- Need to choose parameters
  - Number hidden layers
  - Learning rates
  - Number of steps
  - Number of epochs

- Need to transform textual data to input data

- Training requires a lot of time
Index

1. Introduction
   - Motivation

2. Method
   - Summary
   - Corpus
   - Techniques

3. Experiment
   - Experiment

4. Results
   - Error measure
   - Training Error
   - Test Error

5. Discussion
Experiment

Models

- 2-grams, 3-grams, 4-grams
- RNN
  - 86 input
  - 300 hidden
  - 86 output
  - 50 steps

Datasets

- English wikipedia
- JRC and wikipedia merged
Index

1 Introduction
   • Motivation

2 Method
   • Summary
   • Corpus
   • Techniques

3 Experiment
   • Experiment

4 Results
   • Error measure
   • Training Error
   • Test Error

5 Discussion
Error measure

**Cross-entropy error**

- Cross-entropy
  \[
  H(p, q) = - \sum_x p(x) \log q(x) \quad (1)
  \]
- For each prediction of a sentence
- Then averaged
  \[
  Error = \frac{1}{N} \sum_{i=1}^N H_i(p_i, q_i) \quad (2)
  \]
Writing with RNN

Results

Training Error

Index

1 Introduction
   • Motivation

2 Method
   • Summary
   • Corpus
   • Techniques

3 Experiment
   • Experiment

4 Results
   • Error measure
   • Training Error
   • Test Error

5 Discussion
• From 22 epochs the test error starts increasing
• Because of the available time we apply one epoch
Index

1 Introduction
   • Motivation

2 Method
   • Summary
   • Corpus
   • Techniques

3 Experiment
   • Experiment

4 Results
   • Error measure
   • Training Error
   • Test Error

5 Discussion
• Large values of N needs more training data
• RNN performs better
Writing with RNN

Results

Test Error

Spanish/English models

- Large values of N needs more training data
- RNN performs better
Discussion

**N-grams**
- Depends on the N size
- Small N do not have a context
- Large N needs more data

**RNN**
- Need more time to train
- Fast in generation time
Unsupervised Learning of Multiple Languages Using Recurrent Neural Networks

Miquel Perelló Nieto, Mathias Berglund¹ and Tapani Raiko¹

Course:
T-61.5910 Research Project in Computer and Information Science

Aalto, Nov 2013