Unsupervised Learning of MultipleLanguages 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

Aalto University
School of Science

Index

1. Introduction
   - Motivation
2. Method
   - Summary
   - Corpus
   - Techniques
3. Experiment
   - Experiment
4. Results
   - Error measure
   - Training Error
   - Test Error
5. Discusion

Aalto University
School of Science

Index

1. Introduction
   - Motivation
2. Method
   - Summary
   - Corpus
   - Techniques
3. Experiment
   - Experiment
4. Results
   - Error measure
   - Training Error
   - Test Error
5. Discusion

Aalto University
School of Science

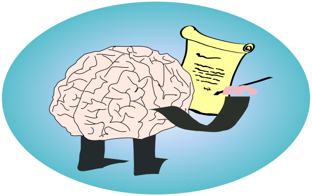
Learning multiple languages

- Le langage est la capacite d'exprimer une pense e et de communiquer au moyen d'un système de signes
- Un idioma ye una llingua, o seya, un sistema de comunicación verbal propiu d'una comunidá 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.

Aalto University
School of Science

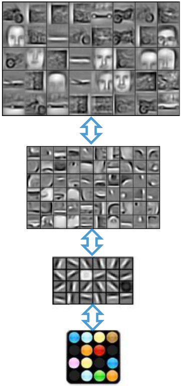
Text prediction

- Involves improving text compression
- Good compresion 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]

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

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

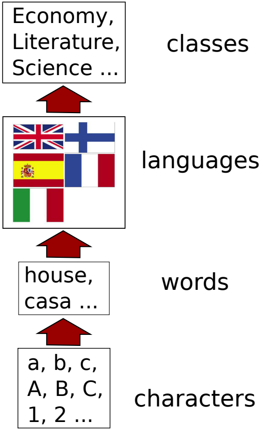
Out: *the tradition of the ancient human reproduction: it is less favorable to the good boy[...]*

[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. Discusion

Summary



Economy, Literature, Science ...

classes

languages

house, casa ...

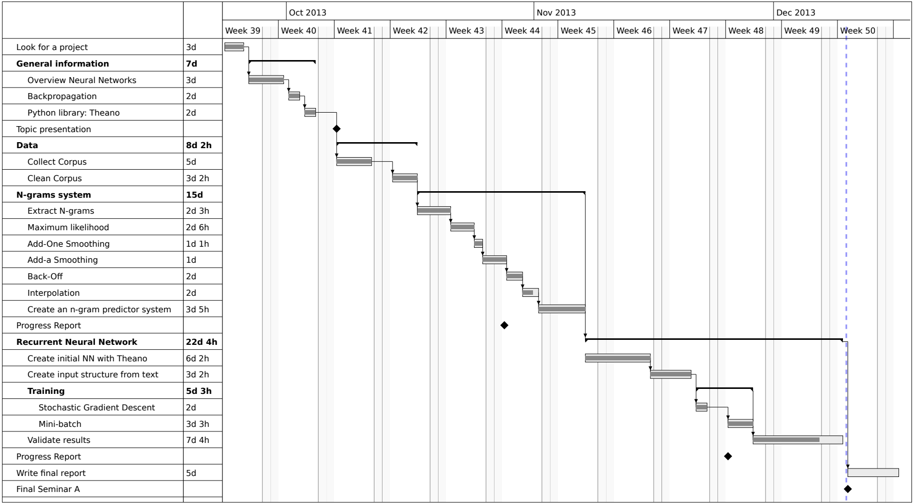
words

a, b, c, A, B, C, 1, 2 ...

characters

- ☑ 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

Timeline



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

Index

Aalto University
School of Science

By language

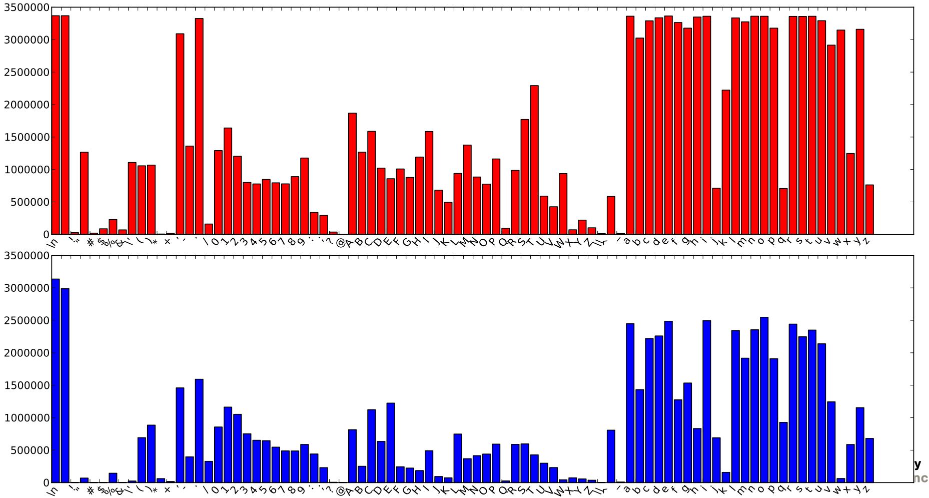
English - 1.4GB

- Wikipedia
- Previously cleaned

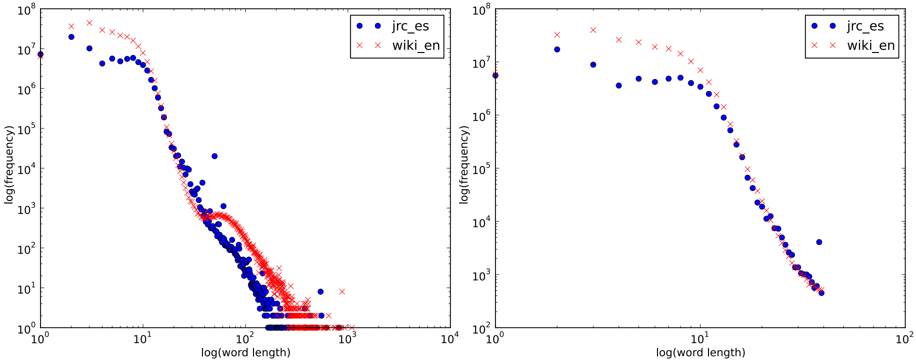
Spanish - 466MB

- 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”

Char frequencies

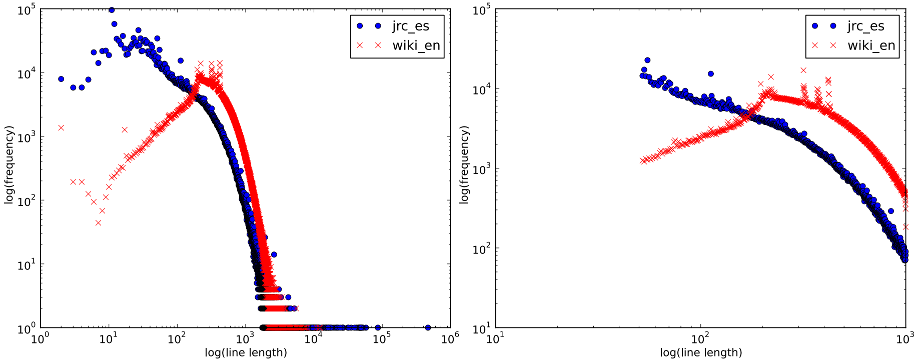


Length words



- 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

Aalto University
School of Science

Index

1. Introduction
   - Motivation
2. Method
   - Summary
   - Corpus
   - Techniques
3. Experiment
   - Experiment
4. Results
   - Error measure
   - Training Error
   - Test Error
5. Discusion

Aalto University
School of Science

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-α 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

Aalto University
School of Science

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

Aalto University
School of Science

Cross-entropy error

- Cross-entropy

$$H(p,q) = -\sum_{x} p(x)\log q(x) \tag{1}$$

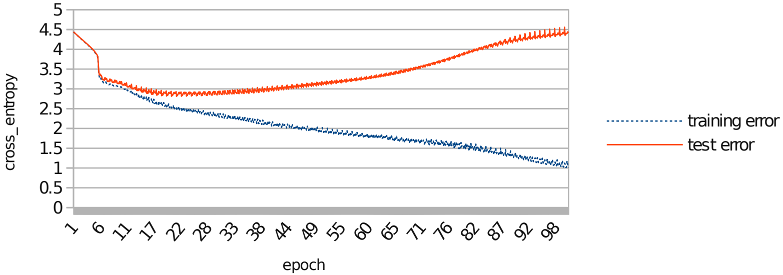
- For each prediction of a sentence
- Then averaged

$$Error = \frac{1}{N}\sum_{i=1}^{N} H_i(p_i, q_i) \tag{2}$$

Index

Aalto University
School of Science

RNN



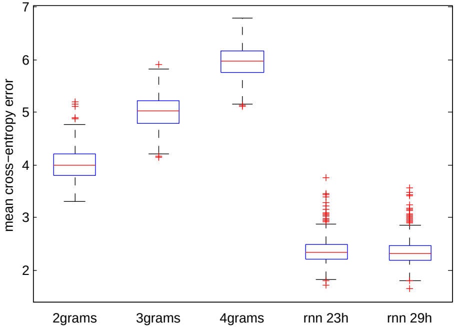
- From 22 epochs the test error starts increasing
- Because of the avaliable time we apply one epoch

Index

1. Introduction
   - Motivation
2. Method
   - Summary
   - Corpus
   - Techniques
3. Experiment
   - Experiment
4. Results
   - Error measure
   - Training Error
   - Test Error
5. Discusion

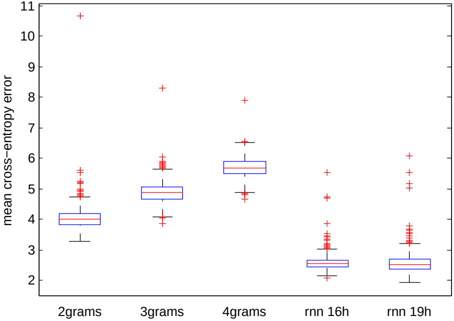
Aalto University
School of Science

English models



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

Spanish/English models



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

Discusion

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

Bibliography I

Ilya Sutskever, James Martens, and Geoffrey E Hinton.
Generating text with recurrent neural networks.
In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.

Unsupervised Learning of MultipleLanguages 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

Aalto University
School of Science