UNIVERSITY OF CAMBRIDGE

Devan Kuleindiren
Robinson College
dk503

Computer Science Tripos - Part II Project

Language Modelling for Text Prediction

March 10, 2017

supervised by
Dr Marek Rei & Dr Ekaterina Shutova

Proforma

Name:	**Devan Kuleindiren**
College:	**Robinson College**
Project Title:	**Language Modelling for Text Prediction**
Examination:	**Computer Science Tripos – Part II, June 2017**
Word Count:	**?**
Project Originator:	Devan Kuleindiren & Dr Marek Rei
Supervisors:	Dr Marek Rei & Dr Ekaterina Shutova

Original Aims of the Project

Work Completed

Special Difficulties

Declaration

I, Devan Kuleindiren of Robinson College, being a candidate for Part II of the Computer Science Tripos, hereby declare that this dissertation and the work described in it are my own work, unaided except as may be specified below, and that the dissertation does not contain material that has already been used to any substantial extent for a comparable purpose.

Signed:

Date:

Contents

List of Figures

Acknowledgements

Chapter 1

Introduction

1.1 Motivation

1.2 Related Work

Chapter 2

Preparation

2.1 N-Gram Models

2.1.1 An Overview of N-Gram Models

Describe how they work, and the motivation for smoothing and backoff.

2.1.2 Smoothing Techniques

Add-One Smoothing

Katz Smoothing

Absolute Discounting

Kneser-Ney

Modified Kneser-Ney

2.2 Recurrent Neural Network Models

2.2.1 An Overview of Neural Networks

Give a brief introduction to neural networks. Explain backpropagation. Motivate why we need RNNs for language modelling.

2.2.2 Recurrent Neural Networks

Explain RNNs.

Vanilla Recurrent Neural Networks

Gated Recurrent Unit

Long Short-Term Memory

2.2.3 Word Embeddings

2.2.4 Backpropagation Through Time

2.3 Software Engineering

2.3.1 Starting Point

2.3.2 Requirements

2.3.3 Tools and Technologies Used

Chapter 3

Implementation

3.1 Development Strategy

3.1.1 Version Control and Build Tools

3.1.2 Testing Strategy

3.2 System Overview

3.2.1 Interface to Language Models

3.3 N-Gram Models

3.3.1 Counting N-Grams Efficiently

3.3.2 Precomputing Smoothing Coefficients

3.4 Recurrent Neural Network Models

3.4.1 The Forward Pass

3.4.2 Gradient Updates

3.4.3 Network Architectures

3.4.4 Parameter Tuning

3.4.5 The Balance Between Underfitting and Overfitting

Dropout, embedding, learning rate decay, momentum?, gradient clipping

3.5 Extending Models to Tackle Error-Prone Text

3.5.1 Preprocessing the CLC Dataset

3.5.2 Error Correction on Word Context

3.6 Benchmarking Framework

3.6.1 Metrics for Accuracy

3.6.2 Metrics for Resource Consumption

3.7 Mobile Keyboard

3.7.1 Updating Language Model Predictions On the Fly

Chapter 4

Evaluation

'Accuracy' could be any one of perplexity, average-keys-saved or guessing entropy.

4.1 Evaluation Methodology

4.2 Results

4.2.1 Existing Models

Things I aim to evaluate here are:
N-Gram Models:

- Accuracy as a function of the amount of training data used (one line per LM), for various smoothing techniques *(on subset of 1BN word dataset)*.
- The effect of increasing N.

Neural Models

- Accuracy as a function of the amount of training data used (one line per LM), for various RNN architectures *(on subset of 1BN word dataset)*.
- The effect of changing the number of hidden neurons.

All Models

- A comparison of the performance of all models *(on the PTB dataset)*.

Combinations of Models

- How combinations of models compare with respect to standalone ones *(on the PTB dataset)*.

4.2.2 On a Mobile Device

Here, I want to focus on the tradeoff between accuracy and resource consumption. Specifically, I could look at the following:

- The effect of increasing the minimum frequency for a word to be considered in the vocabulary. (I.e. the effect of changing the vocabulary size).
- The effect of changing the number of hidden layer neurons.
- The effect of using RNN vs GRU vs LSTM.
- (Perhaps also the effect of pruning on n-gram models).

4.2.3 On Error-Prone Text

Things I aim to evaluate here are:

- The hypothetical upper and lower bounds on accuracy (i.e. the LM results on correct input and on incorrect input respectively).
- The effect of using the vocabulary vs different sized dictionaries for determining if a word should be replaced or not.
- The effect of edit distance on performance.
- An intuitive explanation behind the gap remaining between the current performance and the upper bound. Perhaps some suggestions for future work.

Chapter 5

Conclusion

Appendix A

Project Proposal