Translating Arabic as Low Resource Language using Distribution Representation and Neural Machine Translation Models

by

Ebtesam Almansor

A dissertation submitted in fulfilment of the requirements for the degree
Master of Science (Research) in Computing Sciences

School of Biomedical Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney

August 2018

Acknowledgements

The research in this thesis would not have been possible without the encouragement and support of special people. Firstly, special thanks to my God for helping throughout this research journey. I would like also, to express my sincerest appreciation to my supervisor, Dr. Ahmed Al-Ani who has motivated and supported me generously through this research project. Also I appreciate his patient, kindness and insightful discussion and creative suggestions.

Additionally, special thanks to my beloved father (Hussain) and my much-loved mother (Thagebah), loving husband (Hadi), our new born baby (Hamad), brothers, sisters and friends (Hayat, Alaa, Fatima, Asma, Shima, Wafagah, Nora, Alaa and Aliah) for their support, inspiration, unfaltering belief in my work and confidence in me, and their patience during my time of intense work where without this, this Master degree would never have been completed.

Certificate of Original Authorship

I certify that the work in this thesis has not been previously submitted for a degree nor has it been submitted as a part of the requirements for other degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Production Note:
Signature removed prior to publication.

August 2018
Ebtesam Almansor

Abstract

Rapid growth in social media platforms makes the communication between users easier. According to that, the communication increased the importance of translating human languages. Machine translation technology has been widely used for translating several languages using different approaches such as rule based, statistical machine translation and more recently neural machine translation. The quality of machine translation depends on the availability of parallel datasets. Languages that lack sufficient datasets have posed many challenges related to their processing and analysis. These languages are referred to as low resource languages.

In this research, we mainly focused on low resource languages, particularly Arabic and its dialects. Dialectal Arabic can be treated as non-standard text that is used in Arab social media and need to be translated to their standard forms. In this context, the importance and the focus of machine translation have been increased recently. Unlike English and other languages, translation of Arabic and its dialects have not been thoroughly investigated, where existing attempts were mostly developed based on statistic and rule-based approaches, while neural network approaches have hardly been considered. Therefore, a distribution representation model (embedding model) has been proposed to translate dialectal Arabic to Modern Standard Arabic. As Arabic is a rich morphology language that has different forms of the same words the proposed model can help to capture more linguistic features such as semantic and syntax features without any rules. Another benefit of the proposed model is that it has the capability

to be trained on monolingual datasets instead of parallel datasets. This model was used to translate Egyptian dialect text to Modern Standard Arabic. We also, built a monolingual datasets from available resources and a small parallel dictionary. Different datasets were used to evaluate the performance of the proposed method. This research provides new insight into dialectal Arabic translation.

Recently, there has been increased interest in Neural Machine Translation (NMT). NMT is a deep learning based model that is trained using large parallel datasets with the aim of mapping text from the source language to the target language. While it shows a promising result for high resource translation languages, such as English, low resource languages face challenges using NMT. Therefore, a number of NMT based models have been developed to translate low resource languages, for instance pre-trained models that utilize monolingual datasets. While these models were used on word level and using recurrent neural networks, which have some limitations, we proposed a hybrid model that combines recurrent and convolutional neural networks on character level to translate low resource languages.

Abbreviations

MSA: Modern Standard Arabic

NLP: Natural Language Processing

EGY: Egyptian dialect

MSA: Modern Standard Arabic MT: Machine Translation

EBMT: Example-based Machine Translation Approach

STM: Statistic Machine Translation

NMT: Neural Machine Translation

RNN: Recurrent Neural Network

CNN: Convolutional Neural Network

DAs: Dialects

LM: Language Model

NNLM: Neural Network Language Model

DARPA:Defense Advanced Research Projects Agency

ALPAC:Automatic Language Processing Advisory Committee

HMM: Hidden Markov Model

LSTM: Long Short-Term Memory

GRU: Gate Recurrent Unit

OOV: Out of Vocabulary

BAMA: Buckwalter Arabic Morphological Analyser

LEV and LA : Levantine

GLF: Gulf Arabic

YEM: Yemeni dialect

CALIMAGLF: Gulf Arabic morphological analyzer

MAGEAD: Morphological, analyzer and generator for the Arabic dialects

SAMA: Standard,Arabic morphological analyzer

MADAMIRA: Morphological,Analysis and Disambiguation of Arabic

CODA: Conventional,orthography for dialectal Arabic

TC: Text Classification

ATC: Arabic Text Classification

BOW: Bag-of-words

TF: Term frequency

TF.IDF: Term Frequent. Inverse Document Frequent

CHI: Chi square

BPSO-KNN: Binary Particle Swarm Optimization -K-Nearest-Neighbour

IG: Information Gain

SACM: Cross Validation, Semi-Automatic Categorisation Method

ACM: Automatic Categorisation Method

CBOW: Continuous Bag of Words

seq2seq: Sequence—to—Sequence

BLEU: Bilingual Evaluation Understudy

Table of contents

Abstract **iv**

Abbreviations **vi**

List of figures **xii**

List of tables **xiv**

1 Introduction **1**

1.1 Introduction 1

1.2 Existing Methodological Limitations 4

1.3 Research Question 5

1.4 Research Aim and Objectives 5

1.5 Organization of the Thesis 5

1.6 Publications 6

2 Background and Concept **7**

2.1 Overview 7

2.2 Machine Translation History 7

2.3 Machine Translation Approaches 9

2.3.1 Linguistic Approach 10

2.3.1.1 Transfer-based method 10

2.3.1.2 Interlingua approach 11
2.3.2 Corpus-based approach 12
2.3.2.1 Example Based Machine Translation Approach (EBMT) 12
2.3.2.2 Statistical Machine Translation (SMT) 12
2.3.2.3 Neural machine translation (NMT) 15
2.4 Language Model 16
2.4.1 N-gram language model 17
2.4.2 Neural network language model (NNLM) 18
2.5 Challenges of the Arabic language 20
2.5.1 Arabic script 20
2.5.2 Grammar 21
2.5.3 Morphology 22
2.5.4 Dialectal Applications and Datasets 23
2.6 Arabic Natural Language Processing Tasks 26
2.6.1 Text classification 26
2.6.2 Text Classification Process 27
2.6.3 Datasets 28
2.6.4 Pre-processing 28
2.6.5 Normalization and excluding stop words 29
2.6.6 Stemming 29
2.6.7 Feature Representation 30
2.6.8 Dimensionality Reduction 32
2.7 Summary 33

3 Review of Normalization and Translation Approaches for non-standard text 34

3.1 Overview 34
3.2 Normalization 34
3.2.1 Noisy Channel model 36
3.2.2 Language model 37

3.2.3 Translation model . . . 37
3.3 Translation . . . 38
3.3.1 Rule-based approach . . . 38
3.3.2 Statistical approach . . . 40
3.3.3 Neural machine translation . . . 41
3.3.4 Low Resource languages . . . 43
3.4 Summary . . . 44

4 Translating Dialectal Arabic as a Low Resource Languages using Word Embedding 45
4.1 Overview . . . 45
4.2 Introduction . . . 46
4.3 Dialectal Arabic Language Challenge . . . 47
4.4 Proposed approach . . . 48
4.4.1 Training phase . . . 50
4.4.2 Testing phase . . . 52
4.5 Experiment Setting . . . 52
4.5.1 Building Monolingual Corpus . . . 52
4.5.2 Experiment . . . 54
4.5.3 Results . . . 55
4.6 Summary . . . 59

5 A Hybrid Neural Machine Translation Technique for Translating Low Resource Languages 60
5.1 Overview . . . 60
5.2 Introduction . . . 60
5.3 Sequence to sequence learning . . . 62
5.4 Convolutional sequence to sequence learning . . . 64
5.5 Character level . . . 65
5.6 Proposed model . . . 66

5.7 Experiment and Results 67
5.7.1 Datasets 67
5.7.2 Experiment setting 67
5.7.3 Results 68
5.8 Summary 71

6 Conclusions and Future Work **72**
6.1 Overview 72
6.2 Conclusion 72
6.3 Future work 73

Bibliography **75**

List of figures

1.1 Arabic dialects in different region in the Middle East [1] 3

2.1 Machine translation history. 8
2.2 Neural machine translation. 9
2.3 The structure of MT approaches. 9
2.4 Transfer-based method architecture. 11
2.5 Interlingua method architecture. 11
2.6 An example of word alignment. 14
2.7 Example of an aligned sentence of Arabic-English 15
2.8 Sequence to sequence model 15
2.9 Structure of encoder-decoder for NMT 17
2.10 N-gram language model examples for the sentence (This is a book) . . 18
2.11 Neural network language model architecture [1] 18
2.12 Arabic tashkiil on letters 20
2.13 Statistics of corpus Arabic dialect corpus 26
2.14 Statistics of MSA-ENG and EGY-ENG parallel data 26
2.15 TC process 27

3.1 Example of different non-standard words 35
3.2 Approaches that are used for normalization 35
3.3 Structure of noisy channel model 37

3.4 Machine translation approaches 38
3.5 Character level neural machine translation 42

4.1 Proposed approach 48
4.2 Continuous Bag of Words (CBOW) 49
4.3 Skip-gram model 50
4.4 The average of four-fold-cross-validation of all datasets 56
4.5 Examples of the translation using the proposed model on character and word level 58

5.1 The architecture of the encoder and decoder 63
5.2 Convolutional sequence to sequence learning 65
5.3 Structure of the proposed model 66
5.4 BLEU scores for En-Vi and Ar-En for different sentences length 70

List of tables

2.1 Examples of the difference between the affixes in MSA and DA-EGY . 22
2.2 Dialectal Arabic Tools 24
2.3 Dialectal Arabic Data sets 25
2.4 Arabic Database 28
2.5 Several algorithms can be used to extract features as explained 31

4.1 Constructed corpus 53
4.2 Examples of translation words from the dictionarie 53
4.3 Modern Standard Arabic Datasets 54
4.4 Parallel Datasets 54
4.5 The average accuracy of CBOW for Top@5 and Top@1 55
4.6 The average accuracy of Skip-gram for Top@5 and Top@1 55
4.7 Examples of Semantic relationship 59
4.8 Examples of Syntactic relationship 59

5.1 Experimental datasets size 68
5.2 The best results obtained using the proposed model 69
5.3 Resulting obtained using the proposed model on Arabic-English and English-Vietnamese pairs 69
5.4 Variation in BLEU score using two different sentence lengths 70