N-ARY TREE-CNN FOR ARABIC SENTIMENT ANALYSIS

By

Shimaa Maher Abdallah Baraka

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2020

N-ARY TREE-CNN FOR ARABIC SENTIMENT ANALYSIS

By

Shimaa Maher Abdallah Baraka

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Under the Supervision of

Prof. Dr. Nevin M. Darwish

Professor
Computer Engineering
Faculty of Engineering, Cairo University

Dr. Mona F. Ahmed

Assistant Professor
Computer Engineering
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2020

N-ARY TREE-CNN FOR ARABIC SENTIMENT ANALYSIS

By
Shimaa Maher Abdallah Baraka

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Approved by the
Examining Committee

_____
Prof. Dr. Nevin M. Darwish, Thesis Main Advisor

_____
Prof. Dr. Aly Hassan Fahmy, Internal Examiner

_____
Prof. Dr. Ahmed Abdelwahed Rafea, External Examiner
(Computer Science Professor at the American University in Cairo)

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2020

Engineer's Name: Shimaa Maher Abdallah Baraka
Date of Birth: 14/02/1990.
Nationality: Egyptian
E-mail: Shimaa.baraka@gmail.com
Phone: 01220934596
Address: 2 Melissa Buildings,
Hassan Aflaton St. Nasr City,
Cairo, Egypt

Registration Date: 01/10/2013
Awarding Date: 02/09/2020
Degree: Master of Science
Department: Computer Engineering

Supervisors:

Prof. Dr. Nevin M. Darwish
Dr. Mona F. Ahmed

Examiners:

Prof. Dr. Nevin M. Darwish (Thesis main advisor)
Prof. Dr. Aly Hassan Fahmy (Internal examiner)
Prof. Dr. Ahmed Abdelwahed Rafea (External examiner)
(Computer Science Professor at the American University in Cairo)

Title of Thesis:

N-ary Tree-CNN for Arabic Sentiment Analysis

Key Words:
sentiment analysis; Arabic sentiment analysis; deep learning; sentence representation; recursive neural nets; convolution neural nets; representation learning

Summary:
Distributed document and sentence representation is an essential step in text classification. Several models have been studied to compose sentences into a fixed length representation. Such models range from simple order-insensitive models, like Bag-of-Words, to sequence based models, like RNN. In this thesis we propose an architecture that takes into account the hierarchal nature of the language, by building on binary Recursive Neural Nets, using CNN as an internal representation building block for N-ary trees. The algorithm is applied on Arabic sentiment analysis as an example text classification task and reduces the error rate by up to 15-20% for several standard datasets.

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualifications at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Shimaa Maher Abdallaha Baraka Date: ….. / …. / ….

Signature:

Acknowledgments

I would like to thank my supervisor, Prof. Dr. Niven Darwish, for her support and help throughout the process of obtaining the degree. She was keen on the progress and technical validity of the research and assisted me to accomplish my work.

I would like to thank my supervisor, Dr. Mona Farouk, for her time and technical support during my thesis work. She has provided me with technical advices, guided and mentored me throughout the process.

I would like to thanks my parents and all my family for their endless support and for motivating me.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Definition
CNN	Convolution Neural Network
CV	Computer Vision
DL	Deep Learning
GRU	Gated Recurrent Unit
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MSA	Modern Standard Arabic
MV-RvNN	Matrix Vector Recursive Neural Network
NLP	Natural Language Processing
NN	Neural Network
RNN	Recurrent Neural Network
RNTN	Recursive Neural Tensor Network
RvNN	Recursive Neural Network
SU-RvNN	Syntactically Untied Recursive Neural Network
SVD	Single Value Decomposition
SVM	Support Vector Machine

Abstract

Distributed document and sentence representation is an essential step in text classification. Several models have been studied to compose sentences into a fixed length representation. Such models range from simple order-insensitive models, like Bag-of-Words, to sequence based models, like RNN (Recurrent Neural Networks). Although RNN provides a representation of an arbitrary long sentence and is sensitive to word order, it suffers from a drawback due to the need to have n time steps in order to process a sentence of length n. CNN (Convolutional Neural Network) offers a potentially fully parallel architecture that is order-sensitive within a context defined by the filters. Despite the potential and speed of CNN, it still does not represent how linguistics hierarchy works. Since Recursive Neural Networks (RvNN) are based on parse tree, they are a natural way to represent the hierarchy of the language, where the tree structure acts as a prior to the model. RvNN has the ability to process a sentence in parallel in terms of the level of the tree, potentially reducing the computation steps from n to log n.

Arabic sentence representation and sentiment analysis have witnessed a great development recently, through the development of bigger and better datasets, investigating with deep models, and training a better word representation. Despite this development, tree structures are not yet thoroughly investigated with commonly used Arabic datasets, without the need for tree-annotated datasets. In this thesis, we present a new architecture that builds on the vanilla binary RvNN to make it n-ary by composing the n children nodes through a CNN. In addition, the effect of a linguistically based hierarchy on model power to classify sentiment is illustrated. The proposed architecture outperforms other existing vanilla Deep Learning architectures, like CNN and LSTM, on sentence level sentiment classification, especially short and medium-length sentences, where it lowers the error rate by up to 15-20% for several standard datasets.

Chapter 1 : Introduction

Representation learning of textual data has gained a lot of interest in the last decade. This is, partially, due to its importance as a step in text classification tasks without the need to engineer features around the problem of interest. Sentiment analysis specifically, as an example of text classification tasks, has a great importance in the research field. One main reason is the relying of many industrial institutes on analyzing the rich content existing in social media. In addition, it also serves as a good example of a cluster of NLP (Natural Language Processing) tasks: sentence classification.

1.1. Sentence Representation Learning

Sentence and document representation is an essential step in text classification. Several models have been studied to compose sentences into a fixed length representation. Such models range from simple order-insensitive models, like Bag-of-Words, to sequence based models, like RNN (Recurrent Neural Network). Although RNN provides a representation of arbitrary long sentences and is sensitive to word order, it suffers from a drawback due to the need to have n time steps in order to process a sentence of length n. CNN (Convolution Neural Network) offers a potentially fully parallel architecture that is order-sensitive within a context defined by the filters. Despite of the potential and speed of CNN, it still does not represent how linguistics hierarchy work, which is the main advantage of RvNN (Recursive Neural Network).

1.2. Arabic Sentiment Analysis

Arabic sentence representation and sentiment analysis have witnessed a great development recently. This has been achieved via the development of bigger and better datasets, investigating with deep models, and training a better word representation. Recent development stems mainly from two reasons. The first is the worldwide interest in sentiment analysis, as an important tool in business analysis and industrial domain. The other reason is the latest breakthroughs in the field of NLP, precisely in Latin based languages, which leaves a lot to be investigated and transferred to the Arabic domain. Despite this development, tree structures are not yet thoroughly investigated with commonly used Arabic datasets without the need for tree-annotated datasets.

1.3. Research Objective

The focus of the thesis is studying the effect of representing the hierarchical nature of the language via n-ary parse trees. The parse tree can be thought of as a prior in learning an effective sentence representation to use end-to-end in the sentiment analysis problem. This effect is studied without the introduction of extra labels, i.e. without tree-level annotation. The elimination of node annotation allows us to study the possibility of applying the algorithms on existing data, without the need to curate special type

annotation. In addition, it supports studying the impact of hierarchy in isolation, without introducing the classification label of every tree node as part of the learning process. This work introduces an architecture to allow a more generic form of parse trees instead of binary parse trees which are usually used in literature. The experiments are carried on Arabic datasets of various length and size statistics. Comparison to other DL (Deep Learning) and classical NLP techniques is demonstrated. Investigation of parallelization possibilities are also carried on, to cover the areas where RNN were lacking in terms of scaling the sentence length without necessary scaling the computation time.

1.4. Organization of the thesis

The remainder of this thesis is organized as follows: Chapter 2, Deep Learning Background, covers the overall picture and essential details of deep learning theory. Chapter 3, Review of Literature, goes into the details of deep learning algorithms that are concerned with sentence representation and especially recursive ones. This chapter covers, at the end, the part of the literature that apply such algorithms in the Arabic sentiment analysis domain. Chapter 4, Proposed Approach, describes the proposed architecture in details, along with implementation-specific details to achieve parallization. Chapter 5, Experimental Results, covers all the experiments setup, achieved results, and analysis. Chapter 6, draws a conclusion based on the previous chapters and discusses potential future work.