AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer & Systems Engineering

Machine Understanding through Unsupervised Web Semantification

A Thesis submitted in partial fulfillment of the requirements of
Master of Science in Electrical Engineering
Computer & Systems Engineering

by
Michel Naim Naguib Gerguis
Bachelor of Science in Electrical Engineering
Computer & Systems Engineering
Faculty of Engineering, Ain Shams University, 2012

Supervised By
Prof. Dr. Mohamed Watheq Ali Kamel El-Kharashi
Dr. Cherif Ramzy Salama

Cairo, 2017

AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer & Systems Engineering

Machine Understanding through Unsupervised Web Semantification

by

Michel Naim Naguib Gerguis
Bachelor of Science in Electrical Engineering
Computer & Systems Engineering
Faculty of Engineering, Ain Shams University, 2012

Examiners' Committee

Name and affiliation	**Signature**
Prof. Dr. Mohsen Abd-ElRazek Ali Rashwan Professor at Electronics and Communications Engineering Dept. Faculty of Engineering, Cairo University.	.....
Prof. Dr. Hoda Korashy Mohamed Professor at Computer and Systems Engineering Dept. Faculty of Engineering, Ain-Shams University.	.....
Prof. Dr. Mohamed Watheq Ali Kamel El-Kharashi Professor at Computer and Systems Engineering Dept. Faculty of Engineering, Ain-Shams University.	.....

Date: March 2017

Statement

This thesis is submitted as a partial fulfillment of Master of Science in Electrical Engineering, Faculty of Engineering, Ain Shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Michel Naim Naguib Gerguis
Michel Naim

.....

Date: March 2017

Researcher Data

Name: Michel Naim Naguib Gerguis
Date of Birth: 27/04/1990
Place of Birth: Cairo, Egypt
Last academic degree: Bachelor of Science
Field of specialization: Electrical Engineering
University issued the degree: Ain Shams University
Date of issued degree: 2012
Current job: Research Software Development Engineer II, Microsoft Advanced Technology Lab in Cairo

Abstract

Abstract

Fine grained classification is now crucial for machine understanding. This thesis introduces ClassifyWiki, a framework that automatically generates Wikipedia-based text classifiers given a small set of positive training articles. The target level of granularity is left to the consumer, allowing ClassifyWiki to build classifiers for persons, sportspersons, or even footballers. The main goal is to simplify the process of collecting hundreds or thousands of Wikipedia pages with the same entity class, using a set of positive articles with sizes possibly as small as 10 pages. ClassifyWiki learned from many previous classifiers that tackled few entity classes in order to build a generic framework on top of them to tackle any entity class. ClassifyWiki's output is not models for some entity classes but a framework tuned through more than a hundred of experiments to generate models for any given positive articles. To test the framework, we manually tagged a data set of 2500 Wikipedia pages with the finest grained types we can. The data set covers 808 unique classes on different levels of granularity. ClassifyWiki was tested over 103 different entity classes varying in size down to only 5 positive articles.

On our blind set, we report that ClassifyWiki achieved a macro-averaged f1-score of 83% for 13 entity classes on different levels with 96% precision and 74% recall using 50 or more positive articles. For the main classes, ClassifyWiki scored 97% for Person class using 299 training instances, 79% for location using 214 instances, and 65% for Organizations using 82 instance.

Also, we present WikiTrends, a new analytics framework for Wikipedia articles. It adds the temporal/spatial dimensions to Wikipedia articles in order to visualize the extracted information converting the big static encyclopedia to a vibrant one. WikiTrends enables the generation of aggregated views in timelines or maps for any user-defined collection from unstructured text. Data mining techniques were applied to detect the nationality, start and end year of existence, gender, and entity class for around 4.85 million pages. We evaluated our extractors over a random set of 100 manually tagged pages. Heat maps of notable football players' counts over history or dominant occupations in some specific era are samples of summarizing Wikipedia's big data in WikiTrends maps. WikiTrends' timelines can easily illustrate interesting fame battles over history between male/female actors, music genres, or even between American, Italian, and Indian films. Through information visualization and simple configurations, WikiTrends starts a new experience in answering questions through a figure. The framework is designed to be easily extended so different information types through new extractors could be integrated.

Finally, we present ASU system submitted in the COLING W-NUT 2016 Twitter Named Entity Recognition (NER) task. We present an experimental study on applying deep learning to extracting named entities (NEs) from tweets. We built two Long Short-Term Memory (LSTM) models for the task. The first model was built to extract named entities without types while the second model was built to extract and then classify them into 10 fine-grained entity classes. In this effort, we show detailed experimentation results on the effectiveness of word embeddings, brown clusters, part-of-speech (POS) tags, shape features, gazetteers, and local context for the tweet input vector representation to the LSTM model. Also, we present a set of experiments, to better design the network parameters for the Twitter NER task. Our system was ranked the fifth out of ten participants with a final f1-score for the typed classes of 39% and 55% for the non typed ones.

Thesis Summary

Summary

This thesis summarizes our efforts to build three modules in the direction of machine understanding. The first module is a framework to build classifiers given any set of Wikipedia pages in any level of granularity possibly in any size. We named it ClassifyWiki. We tested our framework over more than 100 entity classes using our dataset based on schema.org. ClassifyWiki does not learn some specific classes like all previous systems but, theoretically, it can generate classifiers for any entity class. We report 83% macro-averaged f1-score using 50 positive training instances.

The second module, we present, is WikiTrends. WikiTrends creates a new analytics layer out of a source of semi-structured and unstructured data. WikiTrends can generate any mix of data to present a new understating of the world. Sample analytics reports were generated like assigning each country some unforgettable additions to humanity, the gender battle down to 1000 BC, tracking trending occupations, musical instruments, and film genres, and summarizing the world view in heat maps.

And the last one is ASU, a system submitted in COLING W-NUT workshop in 2016. The system tackled Twitter Named Entity Recognition task. Our system experimentally shows an incremental approach in designing two LSTM models: One for entity detection and the other for extracting and classifying on a set of 10 fine-grained classes. This study presents experimentally the effect of adding/removing many features in the input representation along with an analysis on the network design. We report a 39% f1-score for the typed model on the test set and a 55% for the non typed one bringing ASU to be the fifth system out of ten participants.

The thesis is divided into six chapters as listed below:

Chapter 1: Introduction

The introduction chapter illustrates the motivation, the background, and the road map of this thesis.

Chapter 2: Related Work

We summarize previous efforts in machine understanding, Wikipedia entity classification, and text analytics in this chapter. We also, position our two frameworks in machine understanding as a general task.

Chapter 3: ClassifyWiki

This chapter presents ClassifyWiki, the data set we developed, our long list of experiments, and our comparisons with previous systems.

Chapter 4: WikiTrends

WikiTrends, the analytics framework, shows the core information extractors, location, time, gender, and entity types in which we plugged ClassifyWiki. Then, we show the magic of a mixture of big data and information extraction to generate visual reports. Finally, examples of timelines and maps that can better answer questions than text.

Chapter 5: Twitter Fine-Grained NER

This chapter summarizes our submission in a shared task in Twitter Fine-Grained NERs in COLING 2016 in the Workshop of Nosiy User-Generated Text (WNUT) that was ranked the fifth out of 10 participants. The chapter shows our experimental study to build this deep learning model and our detailed results.

Chapter 6: Conclusion & Future Work

In this chapter, we summarize our efforts, document our contributions, and list many potential directions for future work for the three modules, ClassifyWiki, WikiTrends, and the Twitter NER named ASU.

Key words: big data, data mining, text analytics, data analytics, entity analytics, text classification, entity classification, fine-grained classification, information extraction, Wikipedia, named entity recognition, twitter, text understanding.

Acknowledgment

First of all, i need to thank my God as he don't leave any thing in my details. He always care and participate with me. He once said, whatever they do prospers, and i feel so as he always give me a hand and even before my hand.

Need to say thank you to my father, mother, sister, and brother. They are the main source of all what i achieved and my comfort zone.

All my thanks to my supervisors, Prod. Dr. Mohamed Watheq Ali Kamel El-Kharashi and Dr. Cherif Ramzy Salama, and also all whoever helped me to reach this point in time. All the doctors and teacher assistants, all organizers, workers, and also cleaners; All, Thank you.

Finally, i hope for the beloved country, i am proud to belong to, Egypt, and for my University, Ain-Shams, to be in a better scientific rank. I tried to do my best to present something through my thesis, even if it is so simple. I hope to give many more so God help me to do so.

Michel Naim Naguib Gerguis
Computer & Systems Engineering
Faculty of Engineering
Ain Shams University
Cairo, Egypt
March 2017

Contents

Contents xv

List of Figures xix

List of Tables xxi

List of Abbreviations xxiii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Motivations . . . 2
- 1.3 Contributions . . . 5
- 1.4 Road Map . . . 6

2 Related Work 9
- 2.1 Machine Understanding . . . 10
- 2.2 Wikipedia Entity Classification . . . 13
- 2.3 Text Analytics . . . 15
- 2.4 Summary . . . 17

3 ClassifyWiki 19
- 3.1 Data Set . . . 20
  - 3.1.1 Rationale . . . 20
  - 3.1.2 Data Set Building Process . . . 21
  - 3.1.3 Entity Classes . . . 21
  - 3.1.4 Data Set Analysis . . . 21
- 3.2 Feature Set . . . 23
  - 3.2.1 Local Features . . . 23
  - 3.2.2 Global Features . . . 24
  - 3.2.3 Features Dominance . . . 25
- 3.3 ClassifyWiki Approach . . . 25
  - 3.3.1 ClassifyWiki Modules . . . 26
  - 3.3.2 Feature Contribution . . . 27
  - 3.3.3 Feature Combinations . . . 28
- 3.4 Experiments and Results . . . 28
  - 3.4.1 Previous Techniques . . . 28
  - 3.4.2 ClassifyWiki Baseline . . . 30
  - 3.4.3 Feature Contribution . . . 30