شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

HANAA ALY

Ain Shams University
Faculty of Computer and Information Sciences
Information Systems Department

NAMED ENTITY RECOGNITION FROM BIOMEDICAL TEXT

Thesis submitted as a partial fulfillment of the requirements for the degree of
Master of Science in Computer and Information Sciences

By

Lobna Ahmed Mady
Teaching Assistant at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Dr. Nagwa Lotfy Badr
Dean of Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Yasmine Mohamed Afify
Lecturer, Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Acknowledgment

I would like to express my sincere gratitude and appreciation to my research main supervisor, Prof. Nagwa Badr. She provided me with valuable guidance and advice throughout this research work. It was really a great opportunity for me to gain from her deep knowledge.

I would like to express my appreciation to my research co-supervisor, Dr. Yasmine Afify, for her valuable comments, support, and encouragement throughout this research. It has been a privilege to work under her supervision and to learn from her experience. This work could not be done without her support.

I would like to thank my backbone and great father, Ahmed Mady, for all the guidance and advice through my whole life. Also, I would like to thank my wonderful mother, Nabwia Sabir, for providing the greatest support and help ever for me as well as for always trying to make my life easy, excellent, and happy.

I would also thank my husband and my other half, Dr. Ahmed Elansary, for being supportive during my master's journey and for always being ready to offer any kind of help.

Also, I want to thank my awesome sister, Soha Mady, for being that kind and caring for me and for all the times she tried to cheer me up and continue my work. More importantly, I have to thank my old brother, Mohamed Mady, for providing help and for all the times he pushed me to continue during hard times. Many thanks also should go to Dr. Abdellatif for his valuable assistance that he provided. Last but not the least, a very special thanks goes to the best friends ever Mariam and Nouran for always being beside me and providing support during whole course of my research.

Abstract

Named entity recognition is commonly considered an essential task in natural language processing. It represents a major phase in information extraction methodology to discover the named entities referenced in unstructured text and classify them into predefined class labels. Identifying biomedical entities has been recognized as a challenging task in named entity recognition.

In this thesis, the applicability of using structured support vector machine to classify flat and nested biomedical entities combined with the feature selection techniques to enhance the performance of biomedical named entity recognition has been thoroughly investigated. The proposed approach used a combination of various types of features to explore the classification performance in a combination of structured support vector machine as a machine learning technique. These features include linguistic, morphological, orthographical, context, and word representation features. The experimental results showed that the performance of the proposed approach surpassed that produced from other benchmark approaches in extracting the biomedical entities such as genes, proteins, cell lines, cell types, DNAs and RNAs.

Derived by these promising results, we were motivated to explore the effect of different types of features on structured support vector machine performance in extracting the biomedical entities. This was achieved by applying two types of filter-based feature selection techniques. The Chi-squared and ReliefF feature selection techniques were utilized to reduce the feature vector and evaluate the importance of each feature. The reduced

feature vector was then taken as an input to structured support vector machine to extract the biomedical entities. Experimental results revealed that the overall performance of biomedical named entity recognition increased when using the adopted feature selection techniques and removed the features with the least effect on the classes.

Most studies in the literature ignored the nested entities and focused on the extraction of flat named entities only. However, the nested entities are commonly used in real world biomedical applications due to their ability to represent semantic meaning of the named entity. Therefore, the proposed approach was evaluated on the extraction of the nested biomedical entities and showed very promising results compared to those obtained from the benchmark approaches.

Comprehensive experiments were conducted on three popular datasets based on five evaluation metrics namely: recall, precision, F1-measure, Geometric Mean, and Matthews correlation coefficient. Experimental results revealed that the structured support vector machine achieved better performance compared to different approaches in the literature for extraction of both flat and nested entities.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
BA	Boundary Assembly
BCC-NER	Bidirectional Contextual Clues Named Entity Recognition
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BNER	Biomedical Named Entity Recognition
CNNs	Convolutional Neural Networks
CNN-LSTM	Convolutional Neural Network and Long Short-Term Memory
CRF	Conditional Random Field
DNA	Deoxyribonucleic Acid
FLSTM	Forward Long Short-Term Memory
FN	False Negative
FP	False Positive
G-mean	Geometric Mean
GPRO	Gene and Gene Product
HMMs	Hidden Markov Models
IE	Information Extraction
KeBioLM	Knowledge enhanced Biomedical pretrained Language Model
LSTM	Long Short-Term Memory Networks
MCC	Matthews correlation coefficient
MEMMs	Maximum Entropy Markov Models
ML	Machine Learning
MultiCNN	Multiple Convolutional Neural Networks
NER	Named Entity Recognition
NEs	Named Entities
Nested BNER	Nested Biomedical Named Entity Recognition
NLP	Natural Language Processing

PCA	Principal Component Analysis
POS	Part of Speech
PSO	Particle Swarm Optimization
RNA	Ribonucleic Acid
SPBA	Statistical Principle Based Approach
SSVM	Structured Support Vector Machine
SVMs	Support Vector Machines
TCN	Temporal Convolutional Network
TCN-CRF	Temporal Convolutional Network with a Conditional Random Field
TP	True Positive
UMLS	Unified Medical Language System
WE	Word Embedding
WR	Word Representation

List of Publications

Lobna A. Mady, Yasmine M. Afify, Nagwa L. Badr, “Biomedical Named Entity Recognition Using Structured Support Vector Machine,” International Journal of Computers and Their Applications, Vol. 28, No. 4, pp. 222-229, 2021.

Lobna A. Mady, Yasmine M. Afify, Nagwa L. Badr, “ Enhancing Performance of Biomedical Named Entity Recognition,” Accepted in International Conference on Intelligent Computing and Information Systems, ICICIS 2021, IEEE, 2021.

Lobna A. Mady, Yasmine M. Afify, Nagwa L. Badr, “ Nested Biomedical Named Entity Recognition,” Accepted in International Journal of Intelligent Computing, and Information Sciences, 2021.