Improving Structured Prediction for Named-Entity Recognition

by

Hanieh Poostchimohammadabadi

A thesis submitted in partial fulfillment for

the degree of DOCTOR OF PHILOSOPHY

School of Electrical and Data Engineering

Faculty of Engineering and Information Technology

University of Technology Sydney

JANUARY 2019

This thesis is dedicated to

my *father*.

Acknowledgments

I would like to sincerely thank my supervisor Prof. Massimo Piccardi for his support and encouragement during this project. At many stages in the course of this research project I benefited from his advice, particularly so when exploring new ideas. His positive outlook and confidence in my research inspired me and gave me confidence. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I would like to thank my family:

- my Mom and Dad for their love, endless support and encouragement. His memory will be always with me;
- my siblings, specially my sister who has always advised me in making right decisions and my older brother whose memory will be eternal;
- my husband for encouraging and reassuring me in all moments of despair.

This research has been funded by the Capital Markets Cooperative Research Centre and is supported by Semantic Sciences Pty Ltd.

Hanieh Poostchi
January 2019, Sydney

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Hanieh Poostchimohammadabadi declare that this thesis, is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has no been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

Signed: Signature removed prior to publication.

Date: 21.01.2019

Abstract

Natural language processing aims to provide an understanding of human utterances adequate to automatically answer questions, translate documents or retrieve information based on its meaning. At the foundation of these capabilities are text analysis tasks such as named-entity recognition (NER) which aims to identify all "named entities" in a text such as people, locations, organizations, numerical expressions and others.

Great effort has been devoted to NER since its inception in 1996. However, the investigation has been mostly focused on languages with large amounts of digital resources such as English, German, Dutch and Spanish. Indeed, NER is still a challenging task for the many languages with low (i.e., little, scarce, scattered) digital resources and manually-annotated corpora. To abridge this gap, in the beginning of this thesis we have targeted NER for a language with scarce annotated resources, namely Persian, that is spoken by a population of over a hundred and ten million people world-wide. To this end, we have provided and published the first manually-annotated Persian NER corpus and introduced an initial NER pipeline that leverages a word embedding and a sequential max-margin classifier. The experimental results show that the proposed approach has been capable of achieving promising MUC7 and CoNLL scores while outperforming two alternatives based on a CRF and a simple RNN. Upon the introduction of the BiLSTM-CRF in 2015, we have mode forward our research by exploring combinations of various word embeddings with the BiLSTM-CRF architecture, with the best combination beating our initial results by more than 12 percentage points.

Building on the achievements of the BiLSTM-CRF in NER, in this thesis we introduce the BiLSTM-SSVM, an equivalent neural model where training is performed using a structured hinge loss. The typical loss functions used for evaluating NER are entity-level variants of the F_1 score such as the CoNLL and MUC losses. Unfortunately, the common loss function used for training NER - the cross entropy - is only loosely related to these evaluation losses. For this reason, we propose a training approach for the BiLSTM-CRF that leverages a hinge loss bounding the CoNLL loss from above. In addition, we present a mixed hinge loss that bounds either the CoNLL loss or the Hamming loss based on the density of entity tokens in each sentence. The experimental results over four benchmark languages (English, German, Spanish and Dutch) show that training with the mixed hinge loss has led to small but consistent improvements over the cross entropy across all languages and four different evaluation measures.

Another interesting NLP component that has been covered in this thesis is cluster naming. Cluster naming is the assignment of representative labels to clusters of documents or words. Once assigned, the labels can play an important role in applications such as navigation, search and document classification. However, finding appropriately descriptive labels is still a challenging task. Accordingly, we have proposed various approaches for assigning labels to word clusters by leveraging word embeddings and the synonymy and hypernymy relations in the WordNet lexical ontology. Experiments carried out using the WebAP document dataset show that one of the approaches stands out in the comparison and is capable of selecting labels that are satisfactorily aligned with those chosen by a pool of four, independent human annotators.

Contents

List of Figures

List of Tables