UNIVERSITY
OF TECHNOLOGY
SYDNEY

Improving Low-Resource Named-Entity Recognition and Neural Machine Translation

by Iñigo Jauregi Unanue

Thesis submitted in fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

under the supervision of Prof. Massimo Piccardi

University of Technology Sydney

Faculty of Engineering and Information Technology

JUNE 2020

This thesis is dedicated to

my *family*.

Acknowledgments

I want to say thank you to my supervisor and friend Prof. Massimo Piccardi, for his brilliant guidance during my PhD. He has always trusted in my capacity as a researcher even though I initially had very limited experience in the field. He has also instilled in me passion for the field we are working in. Finally, he has taught me how to persevere in order to pursue good ideas and achieve my goals. Working with him has been an absolute pleasure.

Nire familiari ere eskerrak eman nahi dizkiot. Aita eta amari, beti nire ondoan egon direlako babes eta maitasun amaigabearekin. Bizitzari modu positiboan eta alai ekiten irakatsi didate eta beti babestu dituzte nire erabakiak. Beraiei esker naiz egun naizen pertsona. Iratiri ere, nire arrebari, eskerrak eman nahi dizkiot. Inork eduki dezaken arrebarik onena da, esan dezaket nire lagunik hoberena dela, eta zoragarria dela elkar zein ondo ulertzen garen. Urte hauetan beti kontatu izan dizkiot nire tesiaren gorabeherak eta beti egon da hor, entzuteko eta laguntzeko prest.

I would also like to dedicate a few, but heartfelt, words to my girlfriend, Nadia. She has been next to me every single day of my thesis, helping and supporting me in moments of despair and sharing the moments of happiness and success. I wouldn't have been able to finish this thesis without her. And to her family, specially to her parents, who have treated me as their own from the first time I came to this country.

I don't want to forget about my friend and colleague Ehsan. He was the one that gave me my first industry work opportunity in Australia, and since then, he has been great research advisor.

Finally, I want to mention that this research has been funded by the Rozetta Institute (formerly known as Capital Markets Cooperative Research Center).

Inigo Jauregi Unanue

June 2020, Sydney

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Inigo Jauregi Unanue 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 referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date: 26.06.2020

Abstract

Named-entity Recognition (NER) and machine translation (MT) are two very popular and widespread tasks in natural language processing (NLP). The former aims to identify mentions of pre-defined classes (e.g. person name, location, time...) in text. The latter is more complex, as it involves translating text from a *source* language into a *target* language.

In recent years, both tasks have been dominated by deep neural networks, which have achieved higher accuracy compared to other traditional machine learning models. However, this is not invariably true. Neural networks often require large human-annotated training datasets to learn the tasks and perform optimally. Such datasets are not always available, as annotating data is often time-consuming and expensive. When human-annotated data are scarce (e.g. low-resource languages, very specific domains), deep neural models suffer from the overfitting problem and perform poorly on new, unseen data. In these cases, traditional machine learning models may still outperform neural models.

The focus of this research has been to develop deep learning models that suffer less from overfitting and can generalize better in NER and MT tasks, particularly when they are trained with small labelled datasets. The main findings and contributions of this thesis are the following. First, health-domain word embeddings have been used for health-domain NER tasks such as drug name recognition and clinical concept extraction. The word embeddings have been pretrained over medical domain texts and used as initialization of the input features of a recurrent neural network. Our neural models trained with such embeddings have outperformed previously proposed, traditional machine learning

models over small, dedicated datasets. Second, the first systematic comparison of statistical MT and neural MT models over English-Basque, a low-resource language pair, has been conducted. This has shown that statistical models can perform slightly better than the neural models over the available datasets. Third, we have proposed a novel regularization technique for MT, based on regressing word and sentence embeddings. The regularizer has helped to considerably improve the translation quality of strong neural machine translation baselines. Fourth, we have proposed using reinforcement-style training with discourse rewards to improve the performance of document-level neural machine translation models. The proposed training has helped to improve the discourse properties of the translated documents such as the lexical cohesion and coherence over various low- and high-resource language pairs. Finally, a shared attention mechanism has helped to improve translation accuracy and the interpretability of the models.

Contents

List of Figures

List of Tables