



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.

Signed: Production Note:
 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 moved 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 intro-

duce 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

Abstract	i
1 Introduction	1
1.1 Objectives and Deliverables	5
1.2 Contributions	6
1.3 Publications	7
1.4 Thesis Outline	8
2 Literature Review	11
2.1 Named-Entity Recognition	11
2.1.1 Unsupervised Word Embedding Methods	12
2.1.1.1 Hellinger-PCA	12
2.1.1.2 GloVe	13
2.1.1.3 Word2Vec	14
2.1.1.4 FastText	16
2.1.2 Sequential Classifiers	17
2.1.2.1 Hidden Markov Models	17
2.1.2.2 Conditional Random Fields	19
2.1.2.3 Structural Support Vector Machines	20
2.1.2.4 Surrogate Loss Functions	23
2.1.3 End-to-End Deep (Recurrent) Neural Networks	25
2.1.3.1 Elman-RNN and Jordan-RNN	28

2.1.3.2	Bidirectional Long Short Term Memory Networks	30
2.1.3.3	BiLSTM-CRF	31
2.1.3.4	Language Model Augmented Sequence Taggers	34
2.1.4	NER Evaluation Metrics	37
2.2	Automatic Cluster Naming	39
2.2.1	Building Hypernym-Hyponym Hierarchical Structures	40
3	Persian Named-Entity Recognition with Structural SVM	43
3.1	Introduction	43
3.2	Related Work	44
3.3	The Proposed Approach	45
3.3.1	Word Embedding	46
3.3.2	Classification	46
3.3.2.1	Sequential Labeling	47
3.3.2.2	Structural SVM	48
3.4	Data Collection	49
3.4.1	PersoSentencesCorpus	49
3.4.2	ArmanPersoNERCorpus	50
3.5	Experiments	55
3.6	Conclusion	56
4	BiLSTM-CRF for Persian Named-Entity Recognition	58
4.1	Introduction	58
4.2	Methods	59
4.2.1	Word Embedding	60
4.2.2	The BiLSTM-CRF for Sequential Labelling	60
4.3	Experimental Results	62
4.4	Conclusion	63

5	BiLSTM-SSVM: Training the BiLSTM to Minimize the CoNLL Loss	65
5.1	Introduction	65
5.2	Related Work	67
5.3	Sequential Labeling	69
5.3.1	BiLSTM-CRF (Cross Entropy)	70
5.3.2	BiLSTM-SSVM (Hinge Loss)	71
5.4	Loss-Augmented Inference Under the CoNLL Loss	72
5.4.1	Mixed Hinge	83
5.5	Proof of Optimality for the Loss-Augmented Inference Algorithm	83
5.6	Experiments and Results	84
5.7	Conclusion	89
6	Cluster Naming Using Word Embeddings and WordNet’s Hypernymy	91
6.1	Introduction and Related Work	91
6.2	The Proposed Pipeline	93
6.2.1	Keyword Extraction	93
6.2.2	Hierarchical Clustering of Keywords	94
6.2.3	Cluster Labeling	95
6.3	Experiments and Results	96
6.3.1	Human Annotation and Evaluation	96
6.3.2	Visualization of Keywords and Hypernyms	97
6.3.3	A Detailed Example	98
6.4	Conclusion	100
7	Conclusion	102
	References	104

List of Figures

Figure 1.1	An example sentence annotated with named entities in the IOB2 format.	2
Figure 2.1	The CBOW and the skip-gram model architectures.	15
Figure 2.2	Hidden Markov Model (HMM) of order one.	17
Figure 2.3	A hyperplane separating two classes.	21
Figure 2.4	Comparison of a number of evaluation and surrogate loss functions for the binary case. The horizontal axis maps the difference in the scores of the correct and incorrect labels; the vertical axis maps the loss.	24
Figure 2.5	Multilayer neural network architecture for sequence tagging (reproduced from [Collobert <i>et al.</i> , 2011]).	26
Figure 2.6	Three types neural networks. (left) Feed-forward NN; (middle) Elman-RNN; (right) Jordan-RNN (reproduced from [Mesnil <i>et al.</i> , 2015]).	29
Figure 2.7	A simple RNN structure for the NER task and its compact visualization.	30
Figure 2.8	A single LSTM memory cell.	31
Figure 2.9	A Bidirectional RNN (reproduced from [Graves <i>et al.</i> , 2013]). . .	32
Figure 2.10	A BiLSTM-CRF model.	33
Figure 2.11	A diagram of the BiLSTM-CRF with an auxiliary LSTM layer for character encoding.	34
Figure 2.12	The TagLM architecture (reproduced from [Peters <i>et al.</i> , 2017]). .	36
Figure 2.13	Taxonomy as a hypernym-hyponym hierarchical structure.	41

Figure 3.1	PersoNER workflow.	45
Figure 3.2	Histogram of the sentences' length in ArmanPersoNERCorpus. . .	52
Figure 3.3	Percentage of sentences containing at least one entity and maxi- mum $p = \{1, \dots, 7\}$ entities of each particular named-entity class.	53
Figure 3.4	A snapshot of ArmanPersoNERCorpus.	54
Figure 4.1	A diagram of the BiLSTM-CRF with an example of prediction. . .	61
Figure 5.1	A pseudo-sentence illustrating all the cases of label transitions. . .	74
Figure 5.2	Training epoch of maximum validation accuracy for the different loss functions.	88
Figure 6.1	The proposed cluster labeling pipeline.	93
Figure 6.2	Precision at k ($P@k$).	98
Figure 6.3	Two-dimensional visualization of an example cluster.	101

List of Tables

Table 2.1	Comparison of the state-of-the-art techniques on CoNLL 2003 English NER dataset.	38
Table 3.1	Class percentages in ArmanPersonNERCorpus.	52
Table 3.2	Comparing the size of ArmanPersonNERCorpus as the first manually annotated Persian NER dataset versus popular NER datasets in other languages.	53
Table 3.3	F_1 score comparison between three different classifiers based on MUC7 and CoNLL score functions for NER task on ArmanPersonNERCorpus.	56
Table 4.1	Comparison of Persian NER results with different classifiers and word embeddings.	63
Table 5.1	The compared training objectives.	85
Table 5.2	Comparison of the CoNLL scores with the different loss functions. . .	86
Table 5.3	Comparison of the MUC scores with the different loss functions. . .	87
Table 5.4	Comparison of the segmentation F_1 scores with the different loss functions.	87
Table 5.5	Comparison of the classification F_1 scores with the different loss functions.	88
Table 5.6	Comparison of the CoNLL scores over the English dataset with the different loss functions and ELMo word embeddings.	89

Table 6.1	An example cluster.	100
-----------	-----------------------------	-----