

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

Abstract	i
1 Introduction	1
1.1 Research Contributions	4
1.2 Publications	6
1.3 Thesis Chapters	8
2 Literature Review	11
2.1 Named-Entity Recognition	11
2.1.1 Traditional NER systems	12
2.1.2 Evaluation in NER	14
2.1.2.1 CoNLL-F1	14
2.2 Machine Translation	15
2.2.1 Traditional MT systems	16
2.2.2 Evaluation in MT	17
2.3 Deep Learning for NER and MT	20
2.3.1 Word Embeddings	21
2.3.1.1 Word2vec	21
2.3.1.2 GloVe	23
2.3.1.3 FastText	24
2.3.1.4 Contextual word embeddings	25
2.3.2 Sentence embeddings	26

2.3.3	Recurrent Neural Networks	27
2.3.3.1	Vanilla RNNs	27
2.3.3.2	LSTM	30
2.3.3.3	GRU	31
2.3.3.4	Bidirectional RNNs	32
2.3.4	Deep Sequential Classification	33
2.3.4.1	BiLSTM-CRF	34
2.3.4.2	Sequence-to-Sequence Models	35
2.3.5	Transformer	37
2.4	Low-Resource Deep Learning	41
2.4.1	Early Stopping	42
2.4.2	Dropout	42
2.4.3	Data Augmentation	43
2.4.4	Multi-task learning	44
2.4.5	Sequence-level training	46
2.4.6	Transfer Learning	48

3	Recurrent Neural Networks with Specialized Word Embeddings for Health-Domain Named-Entity Recognition	51
3.1	Introduction	51
3.2	Related work	54
3.3	Methods	56
3.3.1	CRF	56
3.3.2	Bidirectional LSTM and bidirectional LSTM-CRF	56
3.4	Word features	58
3.4.1	Specialized word embeddings	58
3.4.2	Character-level embeddings	59
3.4.3	Feature augmentation	60
3.5	Results	60

3.5.1	Datasets	60
3.5.2	Evaluation metrics	61
3.5.3	Training and hyper-parameters	63
3.5.4	Results	64
3.5.4.1	CCE results over the i2b2/VA dataset	64
3.5.4.2	DNR results over the DrugBank and MedLine datasets	66
3.5.4.3	Accuracy by entity classes	67
3.6	Conclusion	68
4	English-Basque Statistical and Neural Machine Translation	69
4.1	Introduction	69
4.2	The Basque Language	71
4.3	Methods	72
4.3.1	Moses SMT	72
4.3.2	Apertium	72
4.3.3	Google Translate	72
4.3.4	OpenNMT	73
4.4	Experiments	74
4.4.1	Corpora	74
4.4.2	Experimental Settings and Results	75
4.5	Conclusion	79
5	Regressing Word and Sentence Embeddings for Regularization of Neural Machine Translation	82
5.1	Introduction	82
5.2	Related Work	84
5.2.1	Regularization Techniques	84
5.2.2	Word and Sentence Embeddings	86
5.2.3	Unsupervised NMT	87

5.3	The Baseline NMT model	88
5.4	Regressing word and sentence embeddings	90
5.4.1	ReWE	90
5.4.2	ReSE	91
5.5	Experiments	93
5.5.1	Datasets	93
5.5.2	Model Training and Hyper-Parameter Selection	95
5.5.3	Results	97
5.5.4	Understanding ReWE and ReSE	100
5.5.5	Unsupervised NMT	104
5.6	Conclusion	107

**6 Leveraging Discourse Rewards for Document-Level Neural Machine Trans-
lation 109**

6.1	Introduction	109
6.2	Related Work	111
6.2.1	Document-level NMT	111
6.2.2	Discourse evaluation metrics	113
6.2.3	Reinforcement learning in NMT	113
6.3	Baseline Models	114
6.3.1	Sentence-level NMT	114
6.3.2	Hierarchical Attention Network	114
6.4	RISK training with discourse rewards	115
6.4.1	Reward functions	116
6.4.2	Mixed objective	118
6.5	Experiments	118
6.5.1	Datasets and experimental setup	118
6.5.2	Results	121
6.5.2.1	Ablation study	123

6.5.2.2	Translation examples	124
6.6	Conclusion	126
7	A Shared Attention Mechanism for Interpretation of Neural Automatic Post-Editing Systems	128
7.1	Introduction	128
7.2	Related work	130
7.2.1	Attention mechanisms for APE	130
7.3	The proposed model	132
7.4	Experiments	133
7.4.1	Datasets	133
7.4.2	Artificial data	133
7.4.3	Training and hyper-parameters	134
7.4.4	Results	135
7.5	Conclusion	139
8	Conclusion	143
	References	146

List of Figures

Figure 1.1	Performance vs data.	3
Figure 1.2	Multilingual NER system.	4
Figure 2.1	NER example.	12
Figure 2.2	Translation example.	16
Figure 2.3	Multiple translations.	18
Figure 2.4	CBOw and Skip-gram.	22
Figure 2.5	GloVe probability matrix.	23
Figure 2.6	ELMo contextualized embeddings.	25
Figure 2.7	Vanilla RNN.	29
Figure 2.8	Unfolded LSTM network.	30
Figure 2.9	LSTM internal architecture.	31
Figure 2.10	GRU internal architecture	32
Figure 2.11	Bidirectional RNN.	33
Figure 2.12	Encoder-decoder architecture.	36
Figure 2.13	Beam search example.	38
Figure 2.14	Transformer based encoder-decoder.	38
Figure 2.15	Transformer encoder.	39
Figure 2.16	Performance of a model over a training (bleu) and test (orange) sets.	41
Figure 2.17	Full model vs dropout.	43
Figure 2.18	Data augmentation in NMT.	45
Figure 2.19	Two different MTL strategies.	45

Figure 2.20	Feature-based approach with pre-trained LM.	49
Figure 2.21	Fine-tuning approach with pre-trained LM.	49
Figure 3.1	(a) DNR and (b) CCE tasks examples, where ‘B’ (beginning) specifies the start of a named entity, ‘I’ (inside) specifies that the word is part of the same named entity, and ‘O’ (outside) specifies that the word is not part of any predefined class.	52
Figure 3.2	The Bidirectional LSTM-CRF with word-level and character-level word embeddings. In the example, word ‘sulfate’ is assumed to be the 5th word in a sentence and its only entity; ‘ \mathbf{x}_5 ’ represents its word-level embedding (a single embedding for the whole word); ‘ \mathbf{x}_5^* ’ represents its character-level embedding, formed from the concatenation of the last hidden state of the forward and backward passes of a character-level Bidirectional LSTM; ‘ \mathbf{h}_1 ’ - ‘ \mathbf{h}_5 ’ are the hidden states of the main Bidirectional LSTM which become the inputs into a final CRF; eventually, the CRF provides the labeling.	57
Figure 3.3	Concatenation of all the word features, including general domain embeddings (bleu), specialized embeddings (green), character-level embeddings (orange) and handcrafted features (red).	59
Figure 3.4	Description of the hand-crafted features.	60
Figure 3.5	(a) An example of an incorrect tagging in the “strict” evaluation method. (b) An example of a correct tagging in the “strict” evaluation method.	62

Figure 5.1 Baseline NMT model. (Left) The encoder receives the input sentence and generates a context vector \mathbf{c}_j for each decoding step using an attention mechanism. (Right) The decoder generates one-by-one the output vectors \mathbf{p}_j , which represent the probability distribution over the target vocabulary. During training \mathbf{y}_j is a token from the ground truth sentence, but during inference the model uses its own predictions. 86

Figure 5.2 Full model: Baseline + ReWE + ReSE. (Left) The encoder with the attention mechanism generates vectors \mathbf{c}_j in the same way as the baseline system. (Right) The decoder generates one-by-one the output vectors \mathbf{p}_j , which represent the probability distribution over the target vocabulary, and \mathbf{e}_j , which is a continuous word vector. Additionally, the model can also generate another continuous vector, \mathbf{r} , which represents the sentence embedding. 88

Figure 5.3 BLEU scores over the de-en test set for models trained with training sets of different size. 100

Figure 5.4 BLEU scores of three models over the enfr validation set for different λ values: baseline (red), baseline + ReWE (MSE) (green), baseline + ReWE (CEL) (blue). Each point in the graph is an average of 3 independently trained models. 101

Figure 5.5 BLEU scores over the Cs-En dev set of a baseline + ReWE + ReSE model, with λ fixed to 20 and different β values. Each point in the graph is an average of 3 independently trained models. 102

Figure 5.6 Visualization of the \mathbf{s}_j vectors from the decoder for a subset of the cs-en test set. Please refer to Section 5.5.4 for explanations. This figure should be viewed in color. 103

Figure 5.7 Visualization of the \mathbf{s}_j vectors in a smaller neighborhood of the center word. 105

Figure 5.8	BLEU scores over the test set. The reported results are the average of 5 independent runs.. The red line represents the baseline model and the blue line is the baseline + ReWE.	106
Figure 6.1	RISK training. Given the source document, the policy (NMT model) predicts l candidate translations. Then, a reward function is computed for each such translation. For supervised rewards, (e.g., BLEU) the reference translation is required, but not for LC and COH. Finally, the RISK loss is computed using the rewards and the probabilities of the candidate translations, differentiated, and backpropagated for parameter update.	115
Figure 6.2	BLEU, LC and COH scores over the Cs-En validation set at different training iterations.	127
Figure 7.1	An example of perfect correction of an <i>mt</i> sentence.	137
Figure 7.2	Partial improvement of an <i>mt</i> sentence.	140
Figure 7.3	Passing on a correct <i>mt</i> sentence.	141
Figure 7.4	A completely incorrect prediction.	142

List of Tables

Table 3.1	Statistics of the training and test datasets used in the experiments . . .	61
Table 3.2	The hyper-parameters used in the final experiments	63
Table 3.3	Comparison of the results between the different RNN models and the state-of-the-art systems over the CCE and DNR tasks.	65
Table 3.4	Percentage of words initialized with pre-trained embeddings in the train, dev and test of the respective datasets.	67
Table 3.5	Results by class for the B-LSTM-CRF with character-level and cc/mimic embeddings.	68
Table 4.1	The number of samples in the <i>PaCo_EnEu</i> , <i>WMT16_IT</i> and <i>Berriak</i> datasets.	74
Table 4.2	BLEU score of the models over the <i>PaCo_EnEu</i> and <i>Berriak</i> corpora.	76
Table 4.3	BLEU score of the models over the <i>WMT16_IT</i> corpus.	77
Table 4.4	Average of the percentages of bypassed words by all the NMT mod- els in each dataset and each direction.	79
Table 4.5	Example of translations over the <i>PaCo2_EnEU</i> (en→eu) test set.	81
Table 5.1	Approximate number of sentences in the each train, dev and test datasets.	94

Table 5.2	BLEU scores over the En-Fr test set. The reported results are the average of 5 independent runs. (†) means that the differences are statistically significant with respect to the baseline with a p-value < 0.05 over a two-tailed Welch’s t-test.	97
Table 5.3	BLEU scores over the Cs-En test set. The reported results are the average of 5 independent runs. (†) means that the differences are statistically significant with respect to the baseline with a p-value < 0.05 over a two-tailed Welch’s t-test.	97
Table 5.4	BLEU scores over the Eu-En test set. The reported results are the average of 5 independent runs. (†) means that the differences are statistically significant with respect to the baseline with a p-value < 0.05 over a two-tailed Welch’s t-test.	98
Table 5.5	BLEU scores over the De-En test set. The reported results are the average of 5 independent runs. (†) means that the differences are statistically significant with respect to the baseline with a p-value < 0.05 over a two-tailed Welch’s t-test.	98
Table 5.6	Clustering indexes of the LSTM models over the cs-en test set. The reported results are the average of 5 independent runs.	102
Table 5.7	Translation examples. Example 1: Eu-En and Example 2: Cs-En.	108
Table 6.1	The datasets used for the experiments.	118
Table 6.2	Main results. (*) means that the differences are statistically significant with respect to the HAN _{join} baseline with a p-value < 0.05 over a one-tailed Welch’s t-test. LC and COH values that come at the expense of a drop in translation accuracy (e.g. BLEU, F_{BERT}) are highlighted in italics.	122
Table 6.3	Ablation study of the various reward functions over the Zh-En TED talks dataset with RISK(1.0). Undesirable LC and COH values are highlighted in italics.	123

Table 6.4	Translation example. Snippet of a document from the Zh-En TED talks test set.	125
Table 6.5	Translation example. Snippet of a document from the Es-En subtitles test set.	125
Table 7.1	The model and its hyper-parameters.	134
Table 7.2	Results on the WMT17 IT domain English-German APE test set.	136
Table 7.3	Percentage of the decoding steps with marked attention weight on either input (<i>src</i> , <i>mt</i>) or both.	139