Neural Machine Translation with Recurrent Networks*

Extended Abstract†

Max W. Portocarrero‡
National University of Saint Augustine
Arequipa, Peru
mxportocarrero@gmail.com

Richard Yantas§
Catholic University Of St. Paul
Arequipa, Peru
richard.yantas@ucsp.edu.com

ABSTRACT

In recent years, neural machine translation(NMT) has been a wide and open area to study Natural Language Processing. With increasing work, there is still to much to do. NMT is an approach to machine translation that uses a large neural network. It departs from phrases-based statistical approaches that use separately engineerd subcomponents.Neural machine translation (NMT) is not a drastic step beyond what has been traditionally done in statistical machine translation (SMT).

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
Max W. Portocarrero and Richard Yantas. 2017. Neural Machine Translation with Recurrent Networks: Extended Abstract. In *Proceedings of intelligent systems conference (Intelligent Systems'18),* Jose E. Ochoa and Ivan Tupac (Eds.). ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

This means the model does not have to explicitly store gigantic phrase tables and language models as in the case of standard MT; hence, NMT has a small memory footprint. Lastly, implementing NMT decoders is easy unlike the highly intricate decoders in standard MT.

2 NEURAL MACHINE TRANSLATIONS

NMT is an approach to machine translation that uses a large neural network. It departs from phrases-based statistical approaches that use separately engineerd subcomponents.Neural machine translation (NMT) is not a drastic step beyond what has been traditionally done in statistical machine translation (SMT). MNT starts emitting one target word a time as illustrated in. NMT is often a large neural network that is trained in an end-to-end fashion and has the abil-ity to generalize well to very long word sequences.

2.1 Encoder and decoder arquitecture

A neural machine translation system is a neural network that directly models the conditional prob-ability $p(y|x)$ of translating a source sentence,x_1,...,x_n to a target sentence y_1,...,y_n A basic form of NMT consists of two components: (a) an which computes a representations for each source sentence and (b) a decoder which generates one target word at a time and hence decomposes the conditional probability as:

$$\log p(y|x) = \sum_{j=1}^{m} logp(y_j|y < j, s)$$

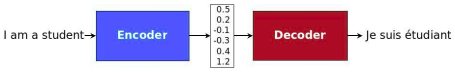

Figura 1. Encoder-decoder architecture

2.2 Long short term memory

We consider as examples a deep multi-layer RNN which is unidirectional and uses LSTM as a recurrent unit. We show an example of such a model in Figure 2. In this example, we build a model to translate a source sentence "I am a student" into a target sentence "Je suis étudiant". At a high level, the NMT model consists of two recurrent neural networks: the encoder RNN simply consumes the input source words without making any prediction; the decoder, on the other hand, processes the target sentence while predicting the next words.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$C_t = tanh(W_c[h_{t-1}, x_t] + b_c)$$

*Produces the permission block, and copyright information
†The full version of the author's guide is available as `acmart.pdf` document
‡Dr. Trovato insisted his name be first.
§The secretary disavows any knowledge of this author's actions.

Intelligent Systems'18, January 2018, Univ. Catolica San Pablo, Arequipa Peru

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

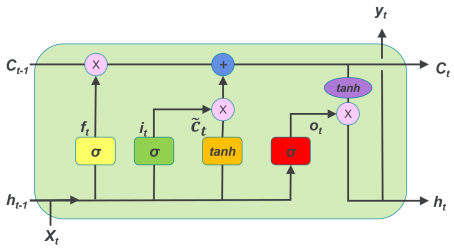

Figura 2. Long short-term memory

3 ATTENTION BASED MODELS

Our various attention-based models are classifed into two broad categories, global and local . These classes differ in terms of whether the attention is placed on all source positions or on only a few source positions.

Common to these two types of models is the fact that at each time step t in the decoding phase, both approaches first take as input the hidden state h_t at the top layer of a stacking LSTM. The goal is then to derive a context vector c_t that captures rel-evant source-side information to help predict the current target word y_t. While these models differ in how the context vector c_t is derived, they share the same subsequent steps.

Specifically, given the target hidden state h_t and the source-side context vector c_t , we employ a simple concatenation layer to combine the infor-mation from both vectors to produce an attentional hidden state as follows:

$$\widetilde{h_t} = tanh(W_c[c_t, h_t])$$

The attentional vector h_t is then fed through the softmax layer to produce the predictive distribu- tion formulated as:

$$p(y_t|y_{<t}, x) = softmax(W_s\widetilde{h_t})$$

3.1 Global attention

The idea of a global attentional model is to con- sider all the hidden states of the encoder when de- riving the context vector c_t. In this model type, a variable-length alignment vector a_t, whose size equals the number of time steps on the source side, is derived by comparing the current target hidden state h_t with each source hidden state h_s.

$$a_t(s) = align(h_t, \bar{h_s})$$
$$= \frac{exp(score(h_t, \bar{h_s}))}{\sum_{s_p} exp(score(h_t, h_s))}$$

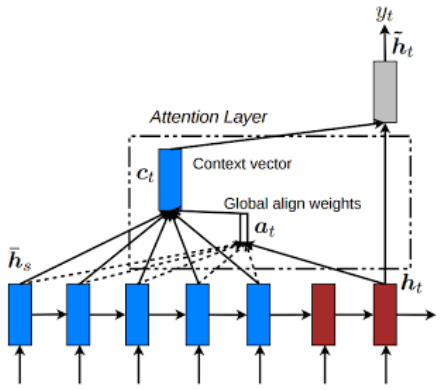

Figura 3. Global attentional model - at each time step t, the models infers a variable-length alignment weigth vector a_t based on the current target state h_t and all source states $\bar{h_s}$. A global context vector c_t is then computed as the weighted average, according to a_t, over all the source states.

Here, score is referred as a *content-based* function for which we consider three different alternatives:

$$score(h_t, \bar{h_s}) = \bar{h_t}^T \bar{h_s}$$
$$score(h_t, \bar{h_s}) = \bar{h_t}^T W_a \bar{h_s}$$
$$score(h_t, \bar{h_s}) = \bar{v_t}^T Tanh(W_a[h_t : \bar{h_s}])$$

Besides alignment vector as weigths, the context vector c_t is computed as the weight average over all the source hidden states.

3.2 Local attention

The global attention has a drawback that it has to attend to all words on the source side for each tar-get word, which is expensive and can potentially render it impractical to translate longer sequences, e.g., paragraphs or documents. To address this deficiency, we propose a local attentional mech-anism that chooses to focus only on a small subset of the source positions per target word.

This model takes inspiration from the tradeoff between the soft and hard attentional models proposed by Xu et al.(2015) to tackle the image caption generation task. In their work, soft attention refers to the global attention approach in which weights are placed "softly" over all patches in the source image. The hard attention, on the order hand, selects on patch of the image to attend to a time.

3.3 Input-feeding Approach

In our proposed global and local approaches, the attentional decisions are made independently, which is suboptimal. Whereas, in standard MT, a coverage set is often maintained during the translation process to keep track of which source words have been translated. Likewise, in atten- tional NMTs, alignment decisions should be made jointly taking into account past alignment information. To address that, we propose an input-feeding approach in which attentional vectors $\bar{h_t}$ are concatenated with inputs at the next time The effects of hav-ing such connections are two-fold: (a) we hope to make the model fully aware of previous align-ment

choices and (b) we create a very deep net- work spanning both horizontally and vertically.

4 EXPERIMENTS

For this Section, we used the base RNN provided by Google's Library Tensorflow with a numerous changes in variables always looking for achiving the best metrics that we could.

Our main metrics were perplexity and Bilingual Evaluation Understudy Score (Bleu) Score. Perplexity is a measurement of how well a probability distribution predicts a sample. It is given by:

$$2^{H(p)} = 2^{-\sum_x p(x) log_2 p(x)}$$

where H(p) is the entropy of the distribution.

So, we may note that the lower the perplexity we get the better our model. But we can not rely on perplexity alone because models can get low perplexity values but that does not mean they get good results. For this, we take into account Bleu Score.

Bleu Score is more about the language and it is strongly correlated with ngrams. This metric is actually an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. The score is calculated for every sentence by compraing them with a set of good quality reference translations. The scores are values between 0 and 1, where 1 means the same referenced sentence. Typically these values are multiplied by 10 (like Google) or 100 (This work).

4.1 Training Details

The dataset used was the European Parliament Parallel Corpus, presented by Philipp Koehn [1]. This dataset is not processed for quick evaluation, it needed some preprocess work. More, we generated our own vocabulary files. These vocabularies were created using 15k Top Frequent 1-grams present in both languages databases with the help of python library Nltk.

In addition, We run NMT Tensorflow code on Intel Core i7-4770 CPU, 8 Gb RAM with a GPU GeForce GTX 750 Ti / 2 Gbs GPU' memory Machine.

Take into account that even with better characteristics training these types of models need several hours or days to complete. We recommend running with tensorflow-GPU configuration and use GPU's with at least 8 Gbs of memory.

Next, we present our different trained models and the values we used.

4.2 English-Spanish Results

On table 1, we present our actual results on our trained models. Next, some sentences translated using our models.

4.3 Sample Translations

On table 3 we can see some translations made by our trained models. In the next section we analize some characteristics of them. Note that special character word <unk> is present when model can not decide what word it is. We assume this happens for a variety of reasons. We can mention lack of vocabulary or not enough training as ones of this reasons.

We only show results on models 3 and 4, which got our best Bleu Scores.

5 ANALYSIS

We compare our results with the ones obtained by article proposed by Luong [2]. The models used by Luong were trained on a dataset were Source language was vietnamite and Target was English. They trained with 12K steps and got a Bleu Score of 25.5.

As we see in Table 2, our best Bleu score is 19.6 in Model 4. Note that this model finish its training it took almost 8 hours training on over 400,000 sentences.

Even though, we note there is still occurrences of the unkwon character '<unk>'. But with increaing training time, we also note its decrease.

Sentences in our samples show their correctness with reference to our target translated sentences. Although, it is not exact, they show the correlation not to words but more surprisingly phrases.

As we can observe, it seems to be a relation between our metrics and training time and vocabulary size, beeing the more important the first one.

6 CONCLUSIONS

Our changes in architecture dropped a series of interesting features on the nmt model used in this work.

We did not mention earlier, but all of our models used the Scaled Luong Attention models. That is mainly because the bleu scores got almost doubled only by activating this characteristic. So we considered to use this addon in our tests.

Traing time is related with a number a features like number of iterations, batch size, number of layers, etc. So, it is important to remember that in order to get the best results the Neural Model needs several hours to train.

We also conclude that NMT is a good translation model. Its potential, scability and no restriction for language pairs make this tecnique very attractive to do more research. Nevertheless, the present goal with the presented method is to accurately find the best training features with may drive in as good metrics scores as the ones gotten by humans.

REFERENCES

[1] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. *http://www.statmt.org/europarl* (2005).

Table 1: Parameters on different Trained Models

Model_Name	Num_Units	Num_Layers	Dropout	Unit_Type	Attention_Model	Train_Steps	Batch
nmt1_en-es	128	2	0.2	lstm	scaled_luong	12,000	128
nmt2_en-es	128	4	0.2	lstm	scaled_luong	40,000	128
nmt3_en-es	128	2	0.2	lstm	scaled_luong	35000	128
nmt4_en-es	128	3	0.2	lstm	scale_luong	100,000	128

Table 2: Perplexity and Bleu Scores on Test Sentences

Model_Name	train_sentences	src_vocab_sz	tgt_vocab_sz	wps	training_time	test_ppl	test_bleu
nmt1_en-es	400,000	15,281	13,306	30.8 K	50 min	18.63	12.6
nmt2_en-es	400,000	15,778	15,488	22.5 K	~ 4h	15.4	13.7
nmt3_en-es	150,000	15,617	14,517	28.2 K	2h 40 min	10.21	15.9
nmt4_en-es	400,000	15,617	14,517	25.8 K	~ 8h	8.23	19.6

Table 3: Sample Translations on Test(unseen) Sentences

Model_Name	Sentences
src:	The proposal by Mrs Malliori is approved and her report will be put to the vote without a debate.
ref:	Queda aprobada la propuesta de la Sra, Malliori y su informe se votará sin debate.
nmt3_en-es	La propuesta de la <unk> <unk> se <unk> y su informe se <unk> a la votación sin un <unk>
nmt4_en-es	La propuesta de la <unk> <unk> se aprueba y su informe se someterá a votación sin un <unk>
src:	The issue is: which exact text will best achieve that? .
ref:	La cuestión es esta: ¿exactamente, qué texto lo logrará mejor? .
nmt3_en-es	La cuestión <unk> que <unk> el texto <unk> <unk> <unk> .
nmt4_en-es	El tema <unk> que el texto exacto se <unk>
src:	Citizens must be able to identify which areas can be regulated by the European Union and which cannot.
ref:	Los ciudadanos deben poder saber qué cuestiones puede regular la Unión Europea y cuáles no.
nmt3_en-es	Los ciudadanos deben poder identificar las zonas que pueden regularse por la Unión Europea y que <unk>
nmt4_en-es	Los ciudadanos deben poder identificar las zonas que pueden regularse por la Unión Europea y que <unk>
src:	That is my main concern in all this, and there are signs that this is again happening already.
ref:	Esa es mi mayor preocupación en este contexto, y hay señales de que esto ya está ocurriendo.
nmt3_en-es	Esta preocupación es mi principal preocupación en todos los <unk> y hay señales que se <unk>
nmt4_en-es	Esta es mi principal preocupación en todos los <unk> y existen indicios de que esto vuelve a ocurrir <unk>
src:	In order to achieve that, it is necessary to promote research, a step the rapporteur also proposes for the Mediterranean in his report.
ref:	Para ello es necesario reforzar la investigación, como también propone el ponente en su texto para el Mediterráneo.
nmt3_en-es	Para alcanzar el <unk> es necesario que se <unk> un paso <unk> el ponente <unk> también el desarrollo del Mediterráneo en su <unk>
nmt4_en-es	Para <unk> es necesario promover el <unk> un paso que propone también el <unk>

[2] Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural Machine Translation (seq2seq) Tutorial. *https://github.com/tensorflow/nmt* (2017).