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, **??** 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:

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

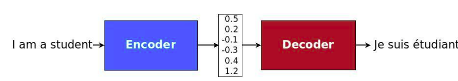

Figura 1. Encoder-decoder architecture

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 Ãľtudiant". 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

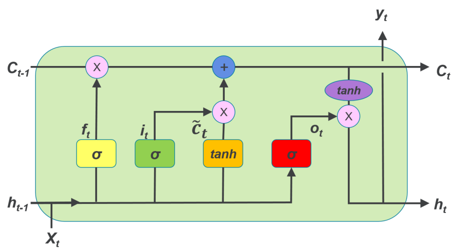

Figura 2. Long short-term memory

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 com-bine 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.

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

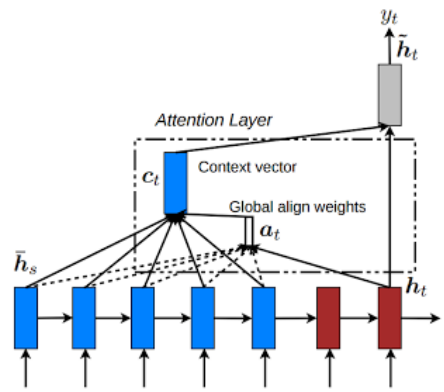

Figura 3. Global attentional model - at each time step t, the models infers a variable-length alignment weigth vector a_t based on the current target state h_t and all source states $\bar{h_s}$. A global context vector c_t is then computed as the weighted average, according to a_t, over all the source 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 deci-sions are made independently, which is suboptimal. Whereas, in standard MT, a coverage set is often maintained during the trans-lation 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 infor-mation. 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.

3.4 Figures

4 EXPERIMENTS

dsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsds dsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsds dsds-dsdsdsdsdsdsvdsdsdsddsdsdsdsdsdsdsdsdsdsds

4.1 Training Details

asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd as-dasdasd

4.2 English-Spanish Results

asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd as-dasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd

4.3 Spanish-English Results

asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd as-dasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd

asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd asdasdasd

5 ANALYSIS

dsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsds dsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsdsds dsds-dsdsdsdsdsdsvdsdsdsddsdsdsdsdsdsdsdsdsdsds

5.1 Learning curves

5.2 Effects of Translations Long Sentences

5.3 Choices of attentional Architectures

5.4 Alignment Quality

5.5 Sample Translations

6 CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

REFERENCES

[] Rafal Ablamowicz and Bertfried Fauser. 2007. CLIFFORD: a Maple 11 Package for Clifford Algebra Computations, version 11. (2007). Retrieved February 28, 2008 from http://math.tntech.edu/rafal/cliff11/index.html

[] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36–44. https://doi.org/10.1145/1188913.1188915

[] American Mathematical Society 2015. *Using the amsthm Package*. American Mathematical Society. http://www.ctan.org/pkg/amsthm.

[] Sten Andler. 1979. Predicate Path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226–236. https://doi.org/10.1145/567752.567774

[] David A. Anisi. 2003. *Optimal Motion Control of a Ground Vehicle*. Master's thesis. Royal Institute of Technology (KTH), Stockholm, Sweden.

[] Mic Bowman, Saumya K. Debray, and Larry L. Peterson. 1993. Reasoning About Naming Systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 795–825. https://doi.org/10.1145/161468.161471

[] Johannes Braams. 1991. Babel, a Multilingual Style-Option System for Use with LaTeX's Standard Document Styles. *TUGboat* 12, 2 (June 1991), 291–301.

[] Malcolm Clark. 1991. Post Congress Tristesse. In *TeX90 Conference Proceedings*. TeX Users Group, 84–89.

[] Kenneth L. Clarkson. 1985. *Algorithms for Closest-Point Problems (Computational Geometry)*. Ph.D. Dissertation. Stanford University, Palo Alto, CA. UMI Order Number: AAT 8506171.

[] Jacques Cohen (Ed.). 1996. Special issue: Digital Libraries. *Commun. ACM* 39, 11 (Nov. 1996).

[] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalances among conjunctive aggregate queries. *J. ACM* 54, 2, Article 5 (April 2007), 50 pages. https://doi.org/10.1145/1219092.1219093

[] Bruce P. Douglass, David Harel, and Mark B. Trakhtenbrot. 1998. Statecarts in use: structured analysis and object-orientation. In *Lectures on Embedded Systems*, Grzegorz Rozenberg and Frits W. Vaandrager (Eds.). Lecture Notes in Computer Science, Vol. 1494. Springer-Verlag, London, 368–394. https://doi.org/10.1007/3-540-65193-4_29

[] Ian Editor (Ed.). 2007. *The title of book one* (1st. ed.). The name of the series one, Vol. 9. University of Chicago Press, Chicago. https://doi.org/10.1007/3-540-09237-4

[] Ian Editor (Ed.). 2008. *The title of book two* (2nd. ed.). University of Chicago Press, Chicago, Chapter 100. https://doi.org/10.1007/3-540-09237-4

[] Simon Fear. 2005. *Publication quality tables in LaTeX*. http://www.ctan.org/pkg/booktabs.

[] Matthew Van Gundy, Davide Balzarotti, and Giovanni Vigna. 2007. Catch me, if you can: Evading network signatures with web-based polymorphic worms. In *Proceedings of the first USENIX workshop on Offensive Technologies (WOOT '07)*. USENIX Association, Berkley, CA, Article 7, 9 pages.

[] David Harel. 1978. *LOGICS of Programs: AXIOMATICS and DESCRIPTIVE POWER*. MIT Research Lab Technical Report TR-200. Massachusetts Institute of Technology, Cambridge, MA.

[] David Harel. 1979. *First-Order Dynamic Logic*. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY. https://doi.org/10.1007/3-540-09237-4

[] Maurice Herlihy. 1993. A Methodology for Implementing Highly Concurrent Data Objects. *ACM Trans. Program. Lang. Syst.* 15, 5 (November 1993), 745–770. https://doi.org/10.1145/161468.161469

[] Lars Hörmander. 1985. *The analysis of linear partial differential operators. III*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. viii+525 pages. Pseudodifferential operators.

[] Lars Hörmander. 1985. *The analysis of linear partial differential operators. IV*. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 275. Springer-Verlag, Berlin, Germany. vii+352 pages. Fourier integral operators.

[] IEEE 2004. IEEE TCSC Executive Committee. In *Proceedings of the IEEE International Conference on Web Services (ICWS '04)*. IEEE Computer Society, Washington, DC, USA, 21–22. https://doi.org/10.1109/ICWS.2004.64

[] Markus Kirschmer and John Voight. 2010. Algorithmic Enumeration of Ideal Classes for Quaternion Orders. *SIAM J. Comput.* 39, 5 (Jan. 2010), 1714–1747. https://doi.org/10.1137/080734467

[] Donald E. Knuth. 1997. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms (3rd. ed.)*. Addison Wesley Longman Publishing Co., Inc.

[] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY.

[] Leslie Lamport. 1986. *LaTeX: A Document Preparation System*. Addison-Wesley, Reading, MA.

[] Newton Lee. 2005. Interview with Bill Kinder: January 13, 2005. Video. *Comput. Entertain.* 3, 1, Article 4 (Jan.-March 2005). https://doi.org/10.1145/1057270.1057278

[] Dave Novak. 2003. Solder man. Video. In *ACM SIGGRAPH 2003 Video Review on Animation theater Program: Part I - Vol. 145 (July 27–27, 2003)*. ACM Press, New York, NY, 4. https://doi.org/99.9999/woot07-S422

[] Barack Obama. 2008. A more perfect union. Video. (5 March 2008). Retrieved March 21, 2008 from http://video.google.com/videoplay?docid=6528042696351994555

[] Poker-Edge.Com. 2006. Stats and Analysis. (March 2006). Retrieved June 7, 2006 from http://www.poker-edge.com/stats.php

[] Bernard Rous. 2008. The Enabling of Digital Libraries. *Digital Libraries* 12, 3, Article 5 (July 2008). To appear.

[] Mehdi Saeedi, Morteza Saheb Zamani, and Mehdi Sedighi. 2010. A library-based synthesis methodology for reversible logic. *Microelectron. J.* 41, 4 (April 2010), 185–194.

[] Mehdi Saeedi, Morteza Saheb Zamani, Mehdi Sedighi, and Zahra Sasanian. 2010. Synthesis of Reversible Circuit Using Cycle-Based Approach. *J. Emerg. Technol. Comput. Syst.* 6, 4 (Dec. 2010).

- [] S.L. Salas and Einar Hille. 1978. *Calculus: One and Several Variable.* John Wiley and Sons, New York.
- [] Joseph Scientist. 2009. The fountain of youth. (Aug. 2009). Patent No. 12345, Filed July 1st., 2008, Issued Aug. 9th., 2009.
- [] Stan W. Smith. 2010. An experiment in bibliographic mark-up: Parsing metadata for XML export. In *Proceedings of the 3rd. annual workshop on Librarians and Computers (LAC '10)*, Reginald N. Smythe and Alexander Noble (Eds.), Vol. 3. Paparazzi Press, Milan Italy, 422–431. https://doi.org/99.9999/woot07-S422
- [] Asad Z. Spector. 1990. Achieving application requirements. In *Distributed Systems* (2nd. ed.), Sape Mullender (Ed.). ACM Press, New York, NY, 19–33. https://doi.org/10.1145/90417.90738
- [] Harry Thornburg. 2001. Introduction to Bayesian Statistics. (March 2001). Retrieved March 2, 2005 from http://ccrma.stanford.edu/~jos/bayes/bayes.html
- [] TUG 2017. Institutional members of the TEX Users Group. (2017). Retrieved May 27, 2017 from http://wwtug.org/instmem.html
- [] Boris Veytsman. [n. d.]. acmart—Class for typesetting publications of ACM. ([n. d.]). Retrieved May 27, 2017 from http://www.ctan.org/pkg/acmart