TACKLING LOW RESOURCE NEURAL MACHINE TRANSLATION APPLIED TO BRAZILIAN PORTUGUESE

Arthur Telles Estrella

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: João Baptista de Oliveira e Souza Filho

Rio de Janeiro
Fevereiro de 2021

TACKLING LOW RESOURCE NEURAL MACHINE TRANSLATION APPLIED TO BRAZILIAN PORTUGUESE

Arthur Telles Estrella

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Orientador: João Baptista de Oliveira e Souza Filho

Aprovada por: Prof. Nome do Primeiro Examinador Sobrenome
Prof. Nome do Segundo Examinador Sobrenome
Prof. Nome do Terceiro Examinador Sobrenome
Prof. Nome do Quarto Examinador Sobrenome
Prof. Nome do Quinto Examinador Sobrenome

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2021

Telles Estrella, Arthur

Tackling low resource neural machine translation applied to Brazilian Portuguese/Arthur Telles Estrella. – Rio de Janeiro: UFRJ/COPPE, 2021.

X, 12 p.: il.; 29, 7cm.

Orientador: João Baptista de Oliveira e Souza Filho

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2021.

Referências Bibliográficas: p. 10 – 11.

1. Primeira palavra-chave. 2. Segunda palavra-chave. 3. Terceira palavra-chave. I. Baptista de Oliveira e Souza Filho, João. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

A alguém cujo valor é digno
desta dedicatória.

Agradecimentos

Gostaria de agradecer a todos.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

TRADUÇÃO DE LINGUAGEM AUTOMÁTICA EM PORTUGUÊS BRASILEIRO ATRAVÉS DE REDES NEURAIS EM DOMÍNIOS DE BAIXO RECURSO

Arthur Telles Estrella

Fevereiro/2021

Orientador: João Baptista de Oliveira e Souza Filho

Programa: Engenharia Elétrica

Apresenta-se nesta dissertação um estudo dedicado a lidar com a tarefa de tradução usando redes neurais, em condições de pouca disponibilidade de dados e com apenas uma GPU, com foco específico para o português-inglês. Será avaliado o efeito prático de técnicas disponíveis na literatura que possuem algum potencial de melhorar a performance nesse contexto, como subword embeddings, pretrained word embeddings e back translation, e como elas impactam qualitativamente no desempenho em frases de diferentes níveis de complexidade. Essas técnicas terão seus prós e contras avaliados e discutidos, utilizando as principais arquiteturas utilizadas na literatura, redes neurais recorrentes e baseadas em transformers. (avaliar isso) O melhor modelo desenvolvido é capaz de atingir x% de score BLEU e y de perplexidade no conjunto de teste do dataset xpto.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TACKLING LOW RESOURCE NEURAL MACHINE TRANSLATION APPLIED TO BRAZILIAN PORTUGUESE

Arthur Telles Estrella

February/2021

Advisor: João Baptista de Oliveira e Souza Filho

Department: Electrical Engineering

In this work, a dedicated analysis is executed to get best practices on how to tackle neural machine translation under low data availability and using a single GPU, focusing specifically on the Portuguese-English pair. Techniques in the literature that can potentially boost the performance under this context will be evaluated, such as subword embeddings, pretrained word embeddings and back translation, and the qualitative impact of them is presented in sentences with a complexity drill down. The tradeoffs of these techniques are discussed, contemplating the main architectures used in the literature, neural recurrent models and transformer-based ones. (evaluate this) The best model built is capable of reaching x% BLEU score and y Perplexity in the test set for the xpto Portuguese dataset.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The reenactment of machine translation

The machine translation is a research field that until 2013 has mainly invested in statistical based models, but the breakthrough promoted by sequence to sequence algorithms followed by the use of transformer models has significantly changed the focus of the field. Before neural networks, machine translation systems were rules-based, syntax-based, phrase-based or a blend between more than one of these techniques. Probabilistic models were used and considered state of the art just before the first sequence to sequence paper appeared. The increase in performance promoted by the sequence to sequence and transformers quickly received some attention, and soon other variants were developed.

Despite being constrained by computational power in many stages since its beginning, one of the most relevant contributions to the translation task were the transformers, which made possible to perform the computation in a truly parallel schema. With this new architecture, the operations performed during training are not totally dependent, allowing them to become parallelizable in the GPU. By removing the constraint of some operations having to wait for others to finish, NMT models were enabled to scale and reach even higher quality translations.

In 3 years NMT became the dominant approach to machine translation, inducing a major transition from statistical to neural models. The broad set of parameters and architectures already present in the literature that could be used to boost translation quality, along with the promising results being presented at the time reached the interest of researchers to explore this variant.

1.2 Challenges for the Portuguese language

Traditionally, the machine translation datasets and conferences usually focus on a subset of languages from countries that are actively investing on NLP, which biases and narrows the potential that the algorithms have towards a specific domain. Unfortunately, Portuguese is a language that does not ostentate supervised translation data in diversity and quantity, an issue that increases the struggle to build a model that can successfully translate it to other languages. Another obstacle is that Portuguese has european, brazilian and african variants, this provides a challenge for a model since generalization is harder if several sentences with different dialects can have the same meaning.

The branch of NMT inside natural language processing is also a field with few papers and academic works among Brazilian universities, this can be partially explained by the challenge that this environment presents: most models require cutting edge GPUs and usually only one GPU is not enough for medium sized model on an average WMT competition dataset. The scarcity of these resources for research purposes require students to innovate in a limited domain and search for cloud solutions without sponsorship.

1.3 Chapter Organization

An explanation of sequence to sequence algorithms and an overview of the most relevant papers regarding NMT and their contributions is given at chapter 2.

In chapter 3, the final model proposed in this work is built piece by piece, where all parameters regarding tuning NMT algorithms are mentioned and the effect of choosing one technique versus another is discussed. This is done via a data-driven approach, the winner techniques are chosen based on BLEU and validation Perplexity results. This chapter leverages the final model potential, finish with a set of most promising parameters to tune.

After reaching a set of potential hyperparameters, chapter 4 contains the final models variants to be tested. These variants are trained on the Portuguese dataset and also on another WMT dataset for comparison purposes.

Finally, in chapter 5 this work is concluded and further improvements and study directions are outlined.

Chapter 2

Neural networks

The application of neural networks has been extended to several domains, at first they were applied to numerical and categorical data, but quickly researchers were able to solve image processing and natural language tasks with some variations on the algorithms. For a long time, this paradigm hasn't seen any applications to solve translation tasks, until Cho et al. [1] who came up with a RNN Encoder-Decoder architecture.

In order to better understand the tradeoffs and capabilities of this technique, we must first have a clear picture of how a recurrent neuron works, which is going to be detailed in the next section.

2.1 Recurrent neural networks

Although being created at the 1980s, recurrent neurons have gained popularity decades later after proving their efficiency through a variety of applications in deep learning.

While feedforward neurons usually struggle to understand relationships in sequential data, the computation behind RNNs depend not only on the current input but also on the current state of the network, which stores information from the past. This turns recurrent neurons into a potential fit to a problem were sequential data is presented.

A vanilla recurrent neuron can be defined by 2 operations, both are updated at each time step t. The first operation maps the input and previous state into the current state, and can be expressed as follows:

$$h^{(t)} = g(W_{xh}^T x^t + W_{hh}^T h^{(t-1)} + b_h) \tag{2.1}$$

Where $h^{(t)}$ is the input state at time step t. The second operation maps the current state into the output, and is given by the equation below:

$$y^{(t)} = W_{hy}^T h^t + b_y \tag{2.2}$$

*** Colocar figurinha dos estados ***

For further reading on how RNNs work and the motivation behind the equations implemented, the reader is referred to SHERSTINSKY [2].

References: https://arxiv.org/pdf/1808.03314.pdf, https://medium.com/@purnasaigudikandula/recurrent-neural-networks-and-lstm-explained-7f51c7f6bbb9, https://classroom.google.com/u/1/c/MTgxMzk1MTY5MTQ4/m/MTg4MT https://builtin.com/data-science/recurrent-neural-networks-and-lstm

2.1.1 Backpropagation through time

Bla

2.1.2 Vanishing and Exploding Gradients

Bla

2.1.3 LSTM

Bla

2.1.4 GRU

Bla

2.2 RNN encoder-decoder architecture

In their approach, one RNN reads each symbol of an input sequence x sequentially and encodes this sequence of symbols into a fixed-length vector representation c, the hidden state. The other decodes this representation into another sequence of symbols by predicting the next symbol y_t given the hidden state h_t. The equation for the decoder's hidden state is given below.

$$h_t = f(h_{t-1}, y_{t-1}, c) \tag{2.3}$$

Bla

2.3 Align and Translate

Bla

2.4 Attention is all you need

Bla

2.5 Recent iterations on seq2seq algorithms

Talk briefly of GPT, BERT and other variations, including using word embeddings

2.5.1 Pretrained word embeddings for NMT

Talk of using Glove, word2vec, fasttext and others in NMT

——- (should I describe them technically?) ——-

Chapter 3

Proposed model

——- (step by step explanation of the hyperparameters explored, results and building piece on piece until we reach out best model) ——-

a

Segundo a norma de formatação de teses e dissertações do Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia (COPPE), toda abreviatura deve ser definida antes de utilizada.

Do mesmo modo, é imprescindível definir os símbolos, tal como o conjunto dos números reais $\mathbb{R}$ e o conjunto vazio $\emptyset$.

> Um exemplo de citação longa nas regras da ABNT (4cm de recuo e fonte menor) feita com o ambiente `longquote` The primary objective of this investigation was to determine the feasibility of detecting corrosion in aluminum Naval aircraft components with neutron radiographic interrogation and the use of standard corrosion penetrameters. Secondary objectives included the determination of the effect of object thickness on image quality, the defining of minimum levels of detectability and a preliminary investigation of a means whereby the degree of corrosion could be quantified with neutron radiographic data. [3]

Para ilustrar a completa adesão ao estilo de citações e listagem de referências bibliográficas, a Tabela 3.1 apresenta citações de alguns dos trabalhos contidos na norma fornecida pela CPGP da COPPE, utilizando o estilo numérico.

Table 3.1: Exemplos de citações utilizando o comando padrão `\cite` do LaTeX e o comando `\citet`, fornecido pelo pacote `natbib`.

Tipo da Publicação	`\cite`	`\citet`
Livro	[4]	ABRAHAM *et al.* [4]
Artigo	[3]	IESAN [3]
Relatório	[5]	MAESTRELLO [5]
Relatório	[6]	GARRET [6]
Anais de Congresso	[7]	GURTIN [7]
Séries	[8]	COWIN [8]
Em Livro	[9]	EDWARDS [9]
Dissertação de mestrado	[10]	TUNTOMO [10]
Tese de doutorado	[11]	PAES JUNIOR [11]

Chapter 4

Results

4.1 Datasets

Apart from the issues outlined, there is a collaborative dataset created and maintained by the Tatoeba Project that contains several English-Portuguese sentence pairs, that is available on the internet. The project contains datasets with several bilingual sentence pairs, considered to match an intermediate english level and translated from English to several other languages, including Portuguese. The Portuguese dataset contains 165k sentences, with 993k words and 21k unique words. It was built by native speakers and has been reviewed by a linguistic expert.

When compared to the average WMT datasets, it is not as large and the vocabulary size is a bit smaller, so it is considered medium level dataset to machine translation, but it suffices our needs.

Chapter 5

Conclusions

References

[1] ET AL, K. C. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation", set. 2014.

[2] SHERSTINSKY, A. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", *CoRR*, v. abs/1808.03314, 2018. Disponível em: <`http://arxiv.org/abs/1808.03314`>.

[3] IESAN, D. "Existence Theorems in the Theory of Mixtures", *Journal of Elasticity*, v. 42, n. 2, pp. 145–163, fev. 1996.

[4] ABRAHAM, R., MARSDEN, J. E., RATIU, T. *Manifolds, Tensor Analysis, and Applications*. 2 ed. New York, Springer-Verlag, 1988.

[5] MAESTRELLO, L. *Two-Point Correlations of Sound Pressure in the Far Field of a Jet: Experiment*. NASA TM X-72835, 1976.

[6] GARRET, D. A. *The Microscopic Detection of Corrosion in Aluminum Aircraft Structures with Thermal Neutron Beams and Film Imaging Methods*. In: Report NBSIR 78-1434, National Bureau of Standards, Washington, D.C., 1977.

[7] GURTIN, M. E. "On the nonlinear theory of elasticity". In: *Proceedings of the International Symposium on Continuum Mechanics and Partial Differential Equations: Contemporary Developments in Continuum Mechanics and Partial Differential Equations*, pp. 237–253, Rio de Janeiro, ago. 1977.

[8] COWIN, S. C. "Adaptive Anisotropy: An Example in Living Bone". In: *Non-Classical Continuum Mechanics*, v. 122, *London Mathematical Society Lecture Note Series*, Cambridge University Press, pp. 174–186, 1987.

[9] EDWARDS, D. K. "Thermal Radiation Measurements". In: Eckert, E. R. G., Goldstein, R. J. (Eds.), *Measurements in Heat Transfer*, 2 ed., cap. 10, New York, USA, Hemisphere Publishing Corporation, 1976.

[10] TUNTOMO, A. *Transport Phenomena in a Small Particle with Internal Radiant Absorption.* Ph.D. dissertation, University of California at Berkeley, Berkeley, California, USA, 1990.

[11] PAES JUNIOR, H. R. *Influência da Espessura da Camada Intrínseca e Energia do Foton na Degradação de Células Solares de Silício Amorfo Hidrogenado.* Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 1994.

Appendix A

Algumas Demonstrações