COPPE
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia
UFRJ

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, 13 p.: il.; 29, 7cm.

Orientador: João Baptista de Oliveira e Souza Filho

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

Referências Bibliográficas: p. 11 – 12.

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.

Finally, Portuguese is a complex language that uses accents which can change their meaning (i.e. "e" and "é"), has different pronoun organizations (i.e. "realizar-se-á" equals "se realizará") and irregular verb inflections (i.e. the "pôr" and "haver" verbs) so text preprocessing and tokenization plays an important role. Disregarding these details by applying some generic preprocessing steps that eliminates accents for instance can lead to worse model performance. On the other hand, having some domain knowledge and apply this to the NLP pipeline can help the model better translate or classify, depending on the desired task.

1.3 Contributions of this dissertation

The generic contribution of this work is the evaluation of a set of techniques available in the literature for NMT that can help dealing with low resource domains, although being applied to Portuguese they can be generalized to other languages. There is also a specialized contribution under a qualitative domain, as performance gains provided by these techniques are evaluated in a subset of sentences that contain a complexity drill down, enabling the reader to have a deeper understanding of their effects on the Portuguese language.

1.4 Chapter Organization

(a ideia final para a dissertação do chapter 2 é falar de arquitetura e revisão teórica/bibliográfica de fundamentos da arquitetura, depois no 3 mostrar restrições que acontecem em low resource domains e suas consequencias, com revisão bibliográfica das técnicas disponíveis que podem ajudar nesse contexto e a motivação das escolhas feitas)

An explanation of the transformer and recurrent neural network architectures and how researchers iterated on them to become state of the art is given in chapter 2.

In chapter 3 constraints that arise in low resource domains and a review of techniques that can potentially help to reduce those issues are presented.

The datasets chosen are presented in chapter 4, where we also describe implementation details, how hyperparameters were iterated on to fit in a single machine GPU and results of the experiments, which are measured in terms of validation perplexity, BLEU and TER(?)(<to be defined>). A qualitative discussion regarding human and model translations in different complexity levels is also performed.

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

Finally, in chapter 5, next steps to conclude this work in the following months are presented

Chapter 2

Neural networks and machine translation

Since their ideation in 1958, neural networks have seen peaks and valleys of research focus in many different artificial intelligence fields. After going out of the tar pit in the 1980s with feedforward and recurrent variants, receiving convolutional and LSTM variants in 1990s, they started to leverage promising results when applied to numerical and categorical data. Specially after 2013 neural networks were growing at an unseen rate, applications to image processing and natural language processing were massively explored and soon consolidated themselves as state of the art algorithms.

For a long time, this paradigm hasn't seen any applications to solve translation tasks, until 2013 when **?**] came up with a RNN Encoder-Decoder architecture proposal. The focus of the scientific community at the time that was on statistical models, but as the application of recurrent networks to this domain gained maturity, the papers gradually started to switch their focus.

2.1 Recurrent neural networks as machine translators

In Cho et al's 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 at time t is given below.

$$h_t = f(h_{t-1}, y_{t-1}, c) \tag{2.1}$$

In this setup, the encoder and decoder of the proposed model are jointly trained to maximize the conditional log-likelihood of a target sequence given a source sequence. It has the drawback of losing some of the information provided by the encoder, since only the hidden state is used, and the RNN output is discarded. Another issue is that the neural network has to compress all the necessary information of a source sentence into a fixed-length vector, which provides a challenge to deal with long sentences.

Massive Exploration of Neural Machine Translation Architectures **?**]

2.1.1 Bahdanau's attention mechanism

NMT by jointly learning to align and translate, comment that it has a new sequence proposal for vectorization

?]

2.1.2 Attention variants for sequence to sequence learning

Effective Approaches to Attention-based Neural Machine Translation LUONG *et al.* [1]

RNN with Doubly-Attentive Decoder **?**]

2.2 Transformer-based models

Attention is all you need **?**]

2.2.1 Have transformers outperformed RNNs?

A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation **?**]

Talk briefly of GPT, maybe bert, and other transformer variations, including using word embeddings

GPT-3 **?**]

DeFINE: DEep Factorized INput Word Embeddings for Neural Sequence **?**]

DeLighT: Deep and Light-weight Transformer **?**]

2.2.2 training techniques can also help

Minimum Risk Training for Neural Machine Translation **?**]

Chapter 3

Low resource context

3.1 low resource techniques

Transfer Learning for Low-Resource Neural Machine Translation **?**]

Revisiting Low-Resource Neural Machine Translation: A Case Study **?**]

Talk of using Glove, word2vec, fasttext and others

Chapter 4

Results

——- (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.

Para ilustrar a completa adesão ao estilo de citações e listagem de referências bibliográficas, a Tabela 4.1 apresenta citações de alguns dos trabalhos contidos na norma fornecida pela CPGP da COPPE, utilizando o estilo numérico.

Table 4.1: Exemplos de citações utilizando o comando padrão `\cite` do LaTeX e o comando `\citet`, fornecido pelo pacote `natbib`.

Tipo da Publicação	`\cite`	`\citet`
Livro	[2]	ABRAHAM *et al.* [2]
Artigo	[3]	IESAN [3]
Relatório	[4]	MAESTRELLO [4]
Relatório	[5]	GARRET [5]
Anais de Congresso	[6]	GURTIN [6]
Séries	[7]	COWIN [7]
Em Livro	[8]	EDWARDS [8]
Dissertação de mestrado	[9]	TUNTOMO [9]
Tese de doutorado	[10]	PAES JUNIOR [10]

Chapter 5

Results

5.1 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 6

Conclusions

References

[1] LUONG, T., PHAM, H., MANNING, C. D. "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, set. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. Disponível em: <`https://www.aclweb.org/anthology/D15-1166`>.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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