Bilingual Knowledge and Ensemble Techniques for Portuguese Natural Language Processing Tasks

Ruan Chaves Rodrigues (UFG) * - ruanchaves93@gmail.com
Jéssica Rodrigues da Silva (B2W Digital) - jsc.rodrigues@gmail.com
Pedro Vitor Quinta de Castro (UFG) * - pedrovitorquinta@inf.ufg.br
Nádia Félix Felipe da Silva (UFG) * - nadia@inf.ufg.br
Anderson da Silva Soares (UFG) * - anderson@inf.ufg.br

* : Institute of Computing
Federal University of Goias (UFG), Brazil

October 15, 2019

DEEP LEARNING BRASIL

Agenda

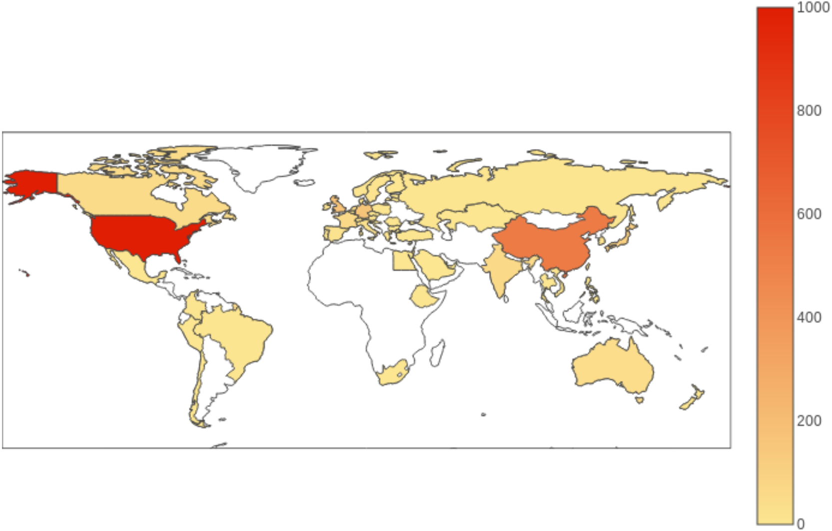
1. Introduction
2. Related Work
3. Ensemble Architecture
4. Experimental Setup
5. Results
6. Conclusions
7. References

Introduction

- **Motivation:**
  - State-of-the-art models are often first trained only in English and/or Chinese.
- **Concept:**
  - Leverage the knowledge of state-of-the-art models through **Automatic Translation**
  - Combine models trained for Portuguese and English through ensemble techniques

Introduction

- **Paper count by country at the 2018 NLP conferences** [Rei 2019]



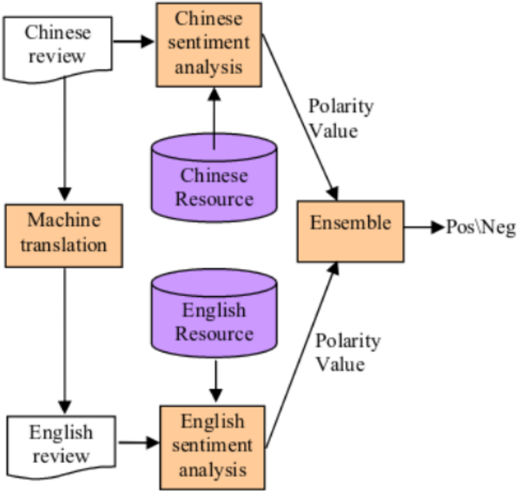
Introduction

- **GLUE Benchmark Leaderboard** [UWNLP 2019]

	Rank	Name	Model	URL	Score
	1	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4
+	2	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.0
	3	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.8
	4	Facebook AI	RoBERTa		88.5
	5	XLNet Team	XLNet-Large (ensemble)		88.4
+	6	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
	7	GLUE Human Baselines	GLUE Human Baselines		87.1

Related Work

- **Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis** [Wan 2008]

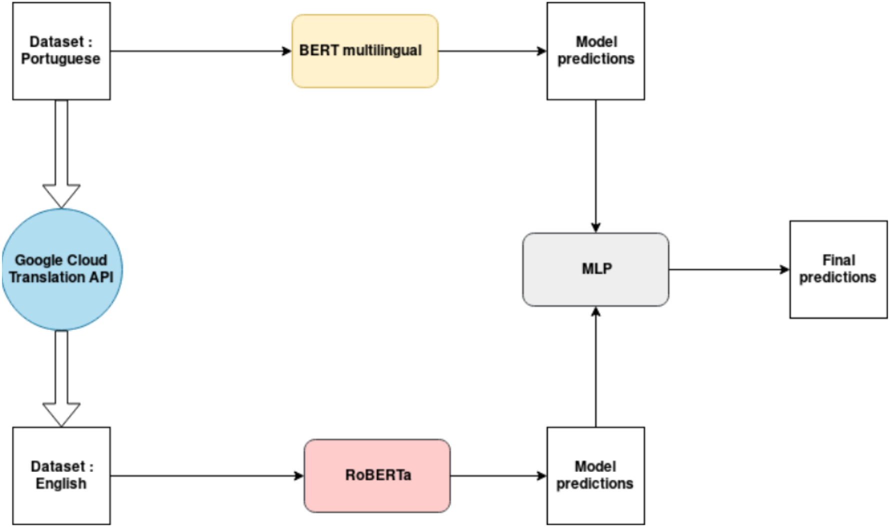


Ensemble Architecture

- **BERT**
  - Bidirectional Encoder Representations from Transformers [Devlin et al. 2018]
- **RoBERTa**
  - "(1) training the model longer, with bigger batches,over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data" [Liu et al. 2019]

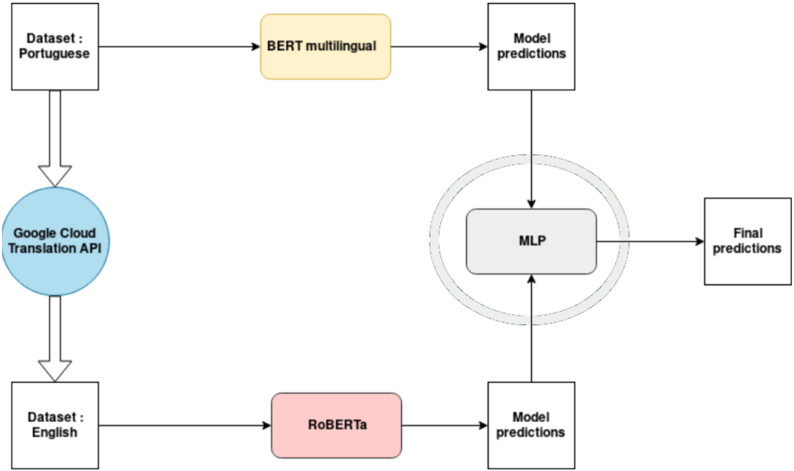
Ensemble Architecture

- **Overview**



Ensemble Architecture

- **Stage I: Training the MLP**

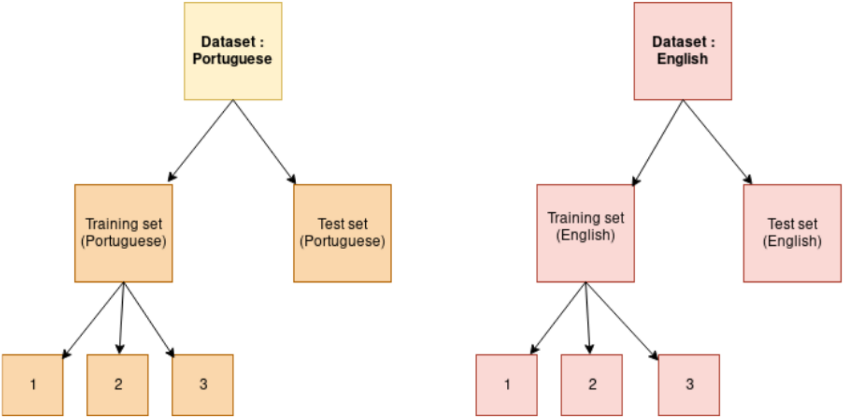


Ensemble Architecture

- **Stage I - Step 1: Translation**

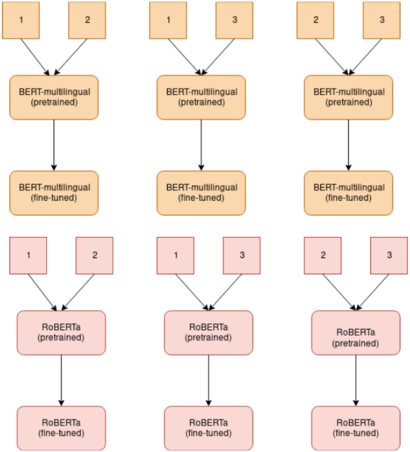
Ensemble Architecture

- **Stage I - Step 2: Split the training data into folds**



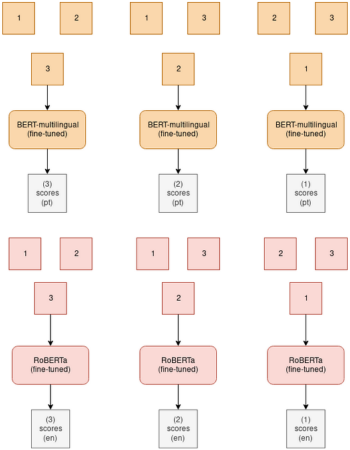
Ensemble Architecture

- **Stage I - Step 3: Fine tune one model for each possible combination of (n - 1) folds**



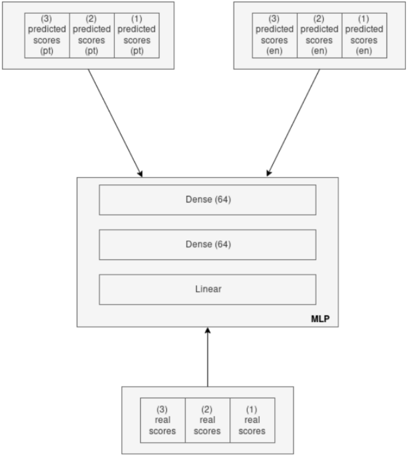
Ensemble Architecture

- **Stage I - Step 4: Predict scores for every missing fold**



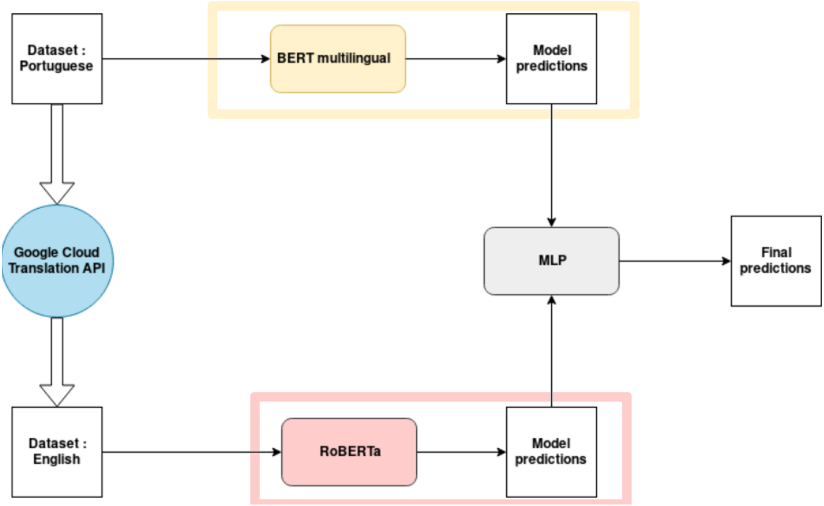
Ensemble Architecture

- **Stage I - Step 5: Train a MLP with the predicted scores**



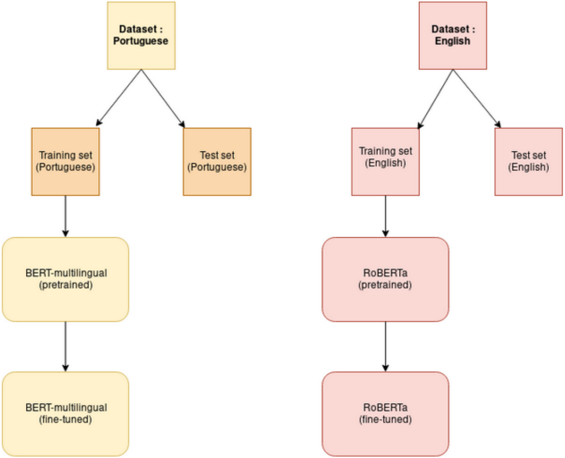
Ensemble Architecture

- **Stage II: Produce the model predictions**



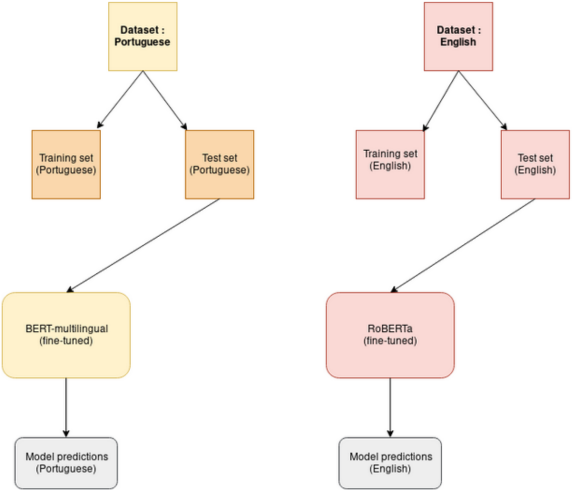
Ensemble Architecture

- **Stage II - Step 6: Fine tune a single model for each training set**



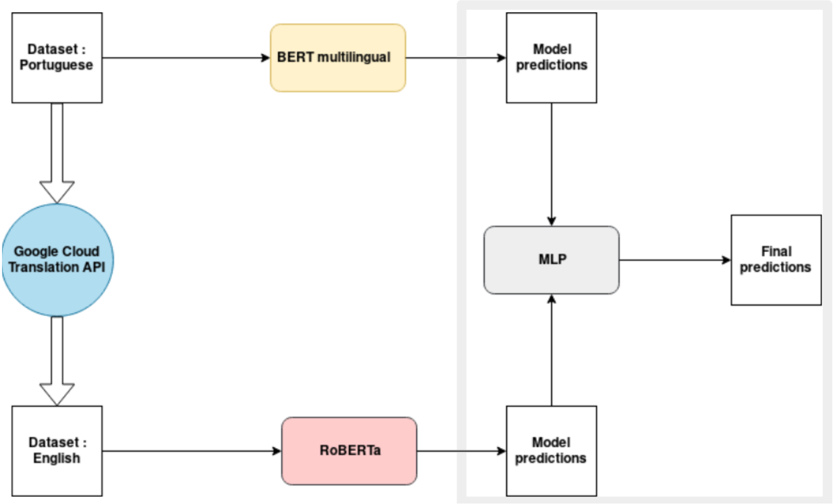
Ensemble Architecture

- **Stage II - Step 7: Let each model predict scores for the test set**



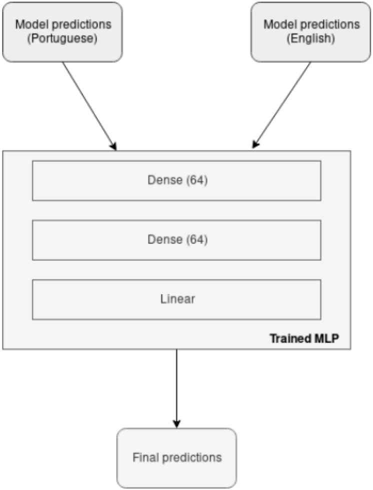
Ensemble Architecture

- **Stage III: Produce the final predictions**



Ensemble Architecture

- **Stage III - Step 8: Combine both model predictions through the now trained MLP**



Ensemble Architecture

- **Stage III - Step 9 (only for entailment): Round up and convert**

[0.345 -0.134 1.128 0.845 … 0.012]

[0 0 1 1 … 0]

[None, None, Entailment, Entailment, …, None]

Experimental Setup

- **Fine-tuning**
  - Transformers library (Hugging Face). [HuggingFace 2019]
  - RoBERTa : 12 epochs.
  - BERT : 4 epochs.
  - Only the final layer is considered.
  - Ensemble architecture: 5 folds
  - The learning rate was adjusted to train the entire ensemble architecture under a single GPU with 8 GB of memory.

Results

Architecture	ASSIN 1 [Propor 2016]		ASSIN 2 [STIL 2019]	
	Entailment	*Similarity*	*Entailment*	*Similarity*
BERT	* * *	0.79	0.819	0.75
RoBERTa	* * *	0.74	**0.884**	**0.81**
Ensemble (5 folds)	* * *	**0.82**	0.883	0.78

ASSIN 1: Complex sentences, most of which will be lost in translation

The ensemble model will always perform better than each model on its own.

ASSIN 2: Simple sentences, which will be almost perfectly translated

The accuracy of the ensemble will approach the stronger model (RoBERTa) as we increase the amount of folds.

Conclusions

- **Highlights**
  - Adaptative ensemble architecture which can produce the best results regardless of the quality of the translation.
  - Robust performance across many domains, thanks to the underlying Transformer architecture.
  - Open source.
- **Future work**
  - Improve on the Multilayer Perceptron.
  - Ensemble BERT and BiMPM. [Wang et al. 2019]
  - Improve BERT's fine-tuning process with SesameBERT. [Su and Cheng 2019]
  - Data augmentation. [Yang et al. 2019]

References I

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

HuggingFace (2019). Transformers. https://huggingface.co/transformers/.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Propor (2016). Assin 1. http://propor2016.di.fc.ul.pt/?page_id=381.

Rei, M. (2019). The geographic diversity of nlp conferences. https://web.archive.org/web/20191009171059/http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/. Accessed: 2019-10-09.

STIL (2019). Assin 2. https://sites.google.com/view/assin2/.

Su, T.-C. and Cheng, H.-C. (2019). Sesamebert: Attention for anywhere.

UWNLP (2019). Glue benchmark leaderboard. https://gluebenchmark.com/leaderboard/. Accessed: 2019-10-09.

Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, R., Su, H., Wang, C., Ji, K., and Ding, J. (2019). To tune or not to tune? how about the best of both worlds?

Yang, W., Xie, Y., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). Data augmentation for bert fine-tuning in open-domain question answering.

Acknowledgements

DEEP LEARNING
BRASIL

INSTITUTO DE
INFORMÁTICA
UFG

Acknowledgements

ANEEL
Agência Nacional de Energia Elétrica
Programa de Pesquisa e Desenvolvimento - P&D

COPEL
Pura Energia

ALBERT EINSTEIN
HOSPITAL ISRAELITA

ifood

AmericasHealth

CYBERLABS

DATA H

B2W
DIGITAL

sixty
LABS
Medicina Exponencial