Tradutor Automatico para Língua Tupi

Rodemarck Júnior[1]**,Tiago Barbosa de Lima**[1] **Andre Nascimento**[1]

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manuel de Medeiros, s/n - Dois Irmãos, Recife - PE, 52171-900 Brazil

{tiago.blima,rodemarck.meloj, andre.camara}@ufrpe.br

Abstract. *Recently neural networks are commonly used to several tasks, like the natural language processing having many applications, and one of them is the translation. Based on that, our project consists in make easier the access to the guarani, language of some Brazilians native speakers, which today has 5 thousand speaks, but it has no automatic translator available, like other native language like the Maori (language spoken for the New Zealand natives).*

Resumo. *Atualmente as redes neurais são comumente utilizados para diversas finalidades, como o processamento de linguagem natural tendo múltiplas aplicações, sendo a tradução uma delas. Baseado nisso, o projeto consiste em facilitar o acesso a guarani, idioma de nativos indígena que atualmente possui 5 mil falantes, porém não possui nenhum tradutor automático como já existem para língua faladas por nativos neozelandenses por exemplo.*

1. Introdução

O Brasil possui muitas riquezas culturais isso inclui a cultura indígena. O Brasil possui 305, dos 826 povos indígenas presentes na América latina [Popolo and Reboiras 2015]. O tupi, ou tupi antigo era a língua falada pelas tribos de povos do grupo tupi, que habitavam a maior parte do litoral brasileiro [Navarro 2013].

O tupi antigo se tornou a língua franca do Brasil Colonia nos seculos XVI e XVII, quando os emigrantes portugueses e seus descendentes aprenderam e o difundiram, mais tarde as expedições bandeirantes levaram, antes restrito aos litorais, para todo o território brasileiro [Navarro 2006].

A língua tupi é sem duvida um de nossos muitos patrimônios culturais, porém é uma língua com baixa acessionabilidade não existindo nenhum tradutor automático para o ela. O desenvolvimento de um tradutor poderia aumentar a acessibilidade a essa língua contribuindo para a preservação e difusão de tal patrimonio.

Na tentativa de criamos um tradutor nós implementamos um algoritmo de *Deep Learning* que podem ser amplamente usados para construção de modelos de inteligência artificial capazes de prever resultados através de textos. Isso possibilita a criação de modelos de tradução de idiomas, tais quais os utilizados pelo *google translator* que não só identifica um determinado idioma, mas também, é capaz de traduzi-lo para muitos outros. No entanto ainda existente vários idiomas que não possuem tradução automática, como é o caso das línguas indígenas brasileiras, a qual são o objeto de estudo do projeto aqui apresentado.

2. Tradução automática

As línguas humanas consistem em morfologia (o modo com que as palavras são montadas a partir de pequenas unidades de sentido), sintaxe (o modo em que as frases são estruturadas) e semântica (o sentido das palavras e/ou frases). A TA (tradução automática) é uma subseção da linguística computacional, é o processo automático de tradução de um idioma original para outro, através do uso de programas de computador. TA não é algo trivial, pois as línguas consistem mais de exceções do que de regras.

3. Tradução por substituição

O método mais simples de implementar um tradutor automático é dividir todas as palavras do texto, traduzir individualmente e substituir pelo texto original, pois tudo o que se precisa é um dicionario.

3.1. Prós

- Muito fácil de implementar.
- Não exige muito processamento.

3.2. Contras

-Confiabilidade extremamente baixa, uma vez que ignora o contexto de cada palavra.
-não leva em consideração a sintaxe da língua.

4. Tradução por regras de sintaxe

Primeiros sistemas de traduções automática, mais sofisticado que a tradução por substituição, pois depois de substituir as palavras por suas traduções aplicavam regras se sintaxe e semântica, para trocar as ordens das palavras adicionar conectivos, quando necessário, e as vezes até substituição da palavra por uma tradução que ofereça um resultado melhor.

4.1. Prós

- Leva em consideração o contexto e as regras de sintaxe
- Pode ter regras adicionadas e atualizadas.
- Tradução coerente e com boa confiabilidade para textos bem formatados e polidos.

4.2. Contras

- Custo extremamente elevado, pois cada língua possui milhares de regras próprias e com ainda mais exceções
- Não consegue bons resultados se houver o menor desvio da linguagem culta.
- Não leva em consideração o uso popular da língua.

5. *Statistical machine translation*

A SMT (*statistical machine translation*, ou tradução estatística de máquina) é uma abordagem diferente, que ao invés de regras de sintaxe usa estatística para traduzir e para isso utiliza uma grande de quantidade de textos possuindo a exata tradução para ambas as linguagens, essa coleção é chamada de corpora paralela. Esta foi a abordagem escolhida para este projeto.

Na SMT o textos não é dividido palavra a palavra, mas em trechos menores, mas que ainda possuam contextos. Em seguida é listado todas as possíveis traduções para cada, não apenas as que estão no dicionario, mas também lista traduções feitas por tradutores reais para aumentar as possibilidades de tradução, cada uma dessa possibilidades ganha uma pontuação baseada na frequência que essa tradução aparece. Milhares de combinações são geradas e a melhor tradução é escolhida baseada em sua pontuação e que seja considerada "mais humana".

5.1. Prós

- Não fica restritos apenas as traduções do dicionário.
- Leva em consideração o uso popular da língua.
- Gera resultados que são mais facilmente entendidos.

5.2. Contras

- Necessita de grande base dados
- Necessita de pré-processamento, pois precisa comparar verso a verso.

6. Materiais

O projeto utiliza palavras coletadas da tradução da Bíblia em Guarani disponível em [Angelo] e uma versão da Bíblia no idioma português Brasil na versão NTLH (Nova Tradução Linguagem de Hoje). A bíblia possui um imenso acervo de palavras (800 mil) se comparamos com outras literaturas similares [Christodouloupoulos and Steedman 2015], além do fato que existem diversas traduções da bíblia segundo a *United Bible Societies* existem 2,527 traduções parciais da bíblia e 475 completas [Christodouloupoulos and Steedman 2015], sendo ideal para o processo de tradução. A bíblia também possui outra característica positiva como sua estrutura com versos bem definidos, sendo ideal para servir como corpora paralela.

Nós utilizamos o algoritmo do *tensorflow "Translate with Attention"* disponível em [ten]. Nós também utilizamos a plataforma do Google colab [goo] para executarmos os teste. Os arquivos foram coletados da bíblia (no formato .csv) do site [Angelo], havendo um preprocessamento em que cada os versos eram alinhados tento em uma coluna com os versos em guarani e na outra em português.

7. Métodos

Após a coleta e preprocessamento dos dados nós executamos 20 *epoch* no nosso algoritmo com *batch* de tamanho 64, dimensão 256 e unidades 1024. Nós começamos com 250 exemplos e aumentamos gradualmente a quantidade de exemplos até 1500 acrescentando a cada interação 250 exemplos usando 80% para treino e 20% para testes.

7.1. Experimentos e Resultados

A imagem da figura 2 é proveniente de um resultado de tradução da seguinte frase: "No começo criou Deus os céus e a terra.".

Nós utilizamos a média do cálculo da distância de Hamming entre a frase traduzida e a tradução literal, considerando (1 - distância de hamming), como sendo a precisão da tradução. Nós variamos os a quantidade de dados de teste obtendo diferentes pontuações para a precisão como mostrado no gráfico da figura 3.

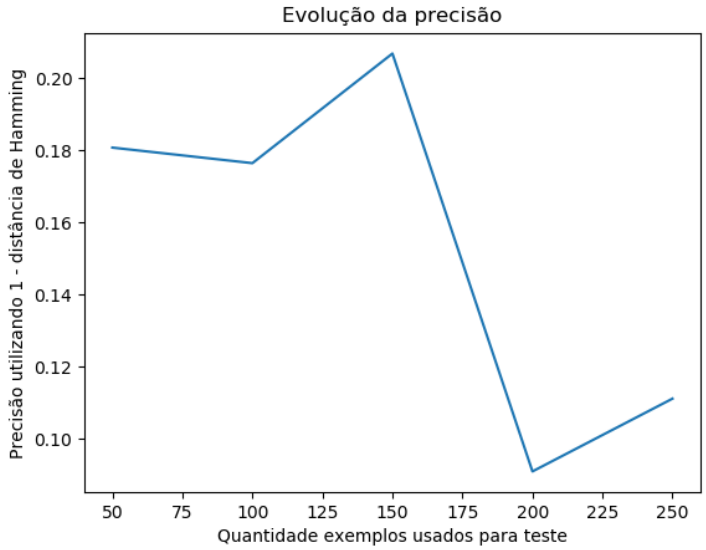

Figure 1. Evolução da acurácia considerando exemplos de 250-1250 variando 250 a cada tentativa.

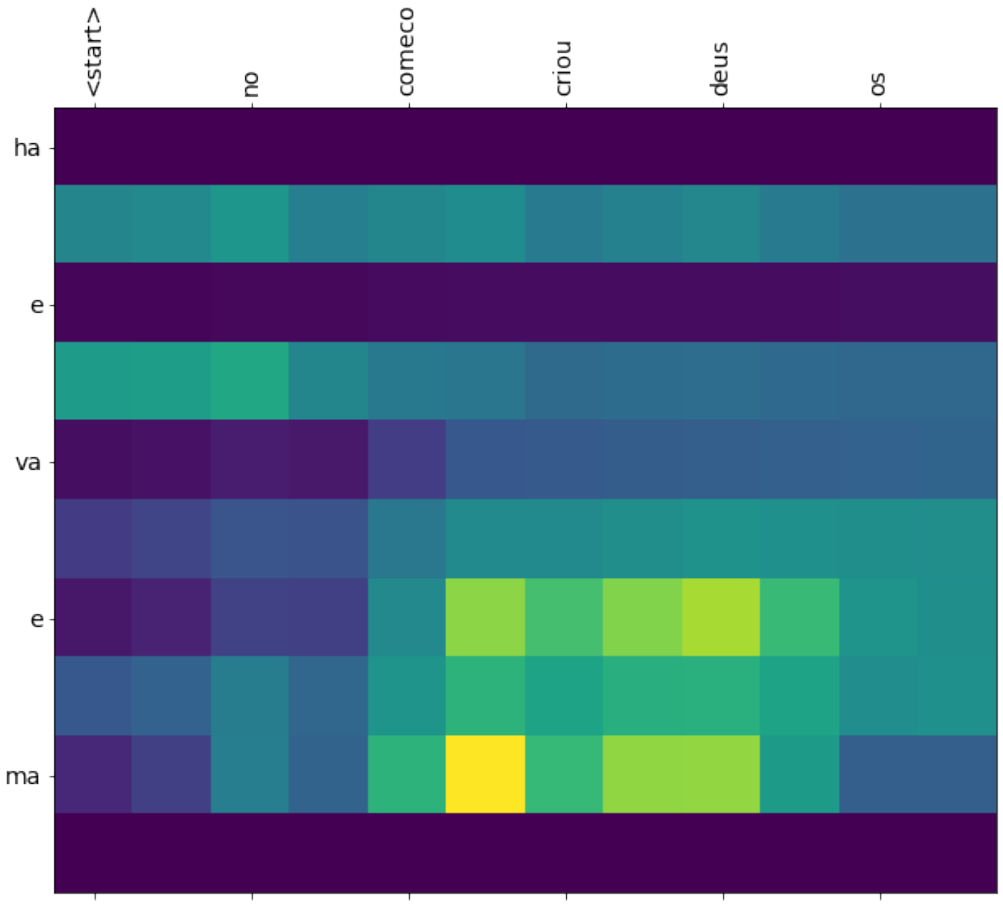

Figure 2. Resultado da tradução na última tentativa considerando 1500 exemplos, é mostrado as palavras levadas em conta com mais atenção na tradução.

8. Conclusão

Nosso tradutor ainda se encontrar no estado inicial precisando alguns modificações e aperfeiçoamentos apesar de já traduzir algumas palavras e frases do guarani para o português brasileiro. Além disso o algoritmo falha ao receber como entrar uma palavra não vista antes o que pode ser melhorado usando técnicas de segmentação de cada palavra. Esperamos com o desenvolvimento desse projeto ajudar na preservação da cultura indígena e no divulgação de idiomas com essa origem.

References

Google colaboratory.

Neural machine translation with attention : Tensorflow core : Tensorflow.

Angelo. 26 versões da bíblia em idiomas indígenas para mysword.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395.

Navarro, E. d. A. (2006). *Método moderno de Tupi antigo - A língua do Brasil dos primeiros séculos*. GLOBAL, 3a edição edition. ISBN 978-85-260-1058-1.

Navarro, E. d. A. (2013). *Dicionário Tupi Antigo - A Língua Índigena Clássica do Brasil*. GLOBAL, 1a edição edition. ISBN 9788526019331.

Popolo, F. and Reboiras, L. (2015). *Os povos indígenas na américa latina: Avanços na última década e desafios pendentes para a garantia de seus direitos*. Nações Unidas.