BRAPT: A New Metric for Translation Evaluation Based on Psycholinguistic Perspectives

R. Guimarães, K. Tavares, R. Reis, L. Ferrari, E. Ogasawara, *Member, IEEE*, and G. Paiva

***Abstract*—There are some metrics to evaluate automatic text translations in the literature. However, the state-of-the-art of these metrics still has limitations. One of them is the dependence of an exact and ordered pairing of words for evaluating similarity among texts. Another, is the non-consideration of the semantics of the text in such comparison. Previous studies point out the need to analyze the semantics of words in the evaluation of translations. In this scenario, this paper presents a novel metric capable of evaluating the differences in automatic text translations that takes into account the semantics of the words presented in the texts. As a proof of concept, we selected ten journalistic texts written in English. These texts have been translated to Portuguese by specialists and by automatic text translation tools. Experimental results show the potential of the proposed metric in evaluating these translations, indicating it can perform better than the state-of-the-art metric.**

***Index Terms*—Automatic text translation, Text mining, BLEU, BRAPT.**

I. Introdução

O PLN (Processamento de Linguagem Natural) compreende um conjunto de técnicas computacionais utilizadas para analisar e representar textos em linguagem natural e tem por propósito auxiliar no processamento de idiomas para o desenvolvimento de diversas tarefas ou aplicações [1]. Dentre as tarefas existentes no PLN, pode-se destacar a geração automática de texto, recuperação da informação e a TAT (Tradução Automática de Textos) [2].

A TAT utiliza recursos do PLN para converter textos de uma linguagem natural para outra, objetivando manter a equivalência semântica entre o conteúdo dos mesmos [3]. Dentre os diversos problemas que podem surgir do PLN, os principais são decorrentes das ambiguidades e não-determinismos [4]. A ambiguidade, por exemplo, pode ser exemplificada com a palavra 'amamos', que pode significar um verbo no presente ou passado. Os não-determinismos são caracterizados por diferentes estruturas na geração da árvore sintática, como por exemplo 'gosto de maçã', que pode gerar duas estruturas distintas, uma considerando 'gosto' como substantivo e outra como verbo.

Nesses aspectos, por utilizar recursos do PLN, as TATs também estão sujeitas a ter sua qualidade comprometida. Dessa maneira, a equivalência semântica no texto traduzido nem sempre é alcançada, podendo gerar traduções que não reflitam o conteúdo original (*e.g.*, apresentando perdas de significado ou de características linguísticas e psicológicas), quando comparadas com a tradução realizada por um especialista humano (*i.e.*, tradução referência) [5]. Como exemplo ilustrador, destaca-se a frase "Ele escreveu que não tinha gado no curral", que submetida à ferramenta Google Tradutor em 01/02/2018, resultou na tradução "He wrote that he had no cattle in the pen".

Nesse cenário, percebe-se que as TATs não são absolutamente fidedignas, sendo necessário que haja um processo de análise/validação (*i.e.*, intervenção humana). No entanto, o processo humano de análise da qualidade de TATs é bastante custoso e pode levar semanas ou meses para ser finalizado [6]. Para lidar com essa problemática, foram propostas algumas métricas para avaliar a qualidade de TATs. Dentre elas, pode-se destacar a métrica BLEU (BiLingual Evaluation Understudy) [7], que ainda permanece como o estado-da-arte [8], sendo a mais utilizada dentre as métricas que avaliam a qualidade das TATs [9].

A ideia central da BLEU é avaliar uma tradução candidata (*i.e.*, realizada por uma ferramenta de TAT) em comparação com uma ou mais traduções referência (*i.e.*, realizada por um especialista humano) [10]. Essa métrica compreende uma média ponderada entre as palavras presentes na tradução referência em comparação com as palavras presentes na tradução candidata. No entanto, métricas desse tipo, por serem baseadas em pareamento exato e ordenado de palavras, não são capazes de considerar similares sentenças diferentes, mas semanticamente semelhantes [11].

Estudos anteriores apontam a necessidade de se analisar a semântica das palavras na avaliação de traduções [3], [12], [13]. Esses estudos destacam a necessidade de não se considerar apenas um pareamento ordenado e exato de palavras nas traduções automáticas, incluindo, por exemplo, características psicolinguísticas das palavras [12]. Um outro exemplo pode ser dado com a inclusão de listas de sinônimos [13].

Nesse panorama, o objetivo principal deste trabalho é propor uma nova métrica (BRAPT) que, diferente das métricas que compõem o estado da arte em avaliação de traduções, compara a tradução referência com a tradução candidata levando em consideração a semântica das palavras presentes nas sentenças. Desse modo, é capaz de envolver informação semântica ao

R. G. Rodrigues, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, RJ - Brasil, rafael.rodrigues@cefet-rj.br.

K. T. Rodrigues, Universidade Federal Fluminense (UFF/RJ), Rio de Janeiro, RJ - Brasil, kaio_rodrigues@id.uff.br.

R. R. Gomes, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, RJ - Brasil, rodrigo.gomes@cefet-rj.br.

L. Ferrari, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ - Brasil, lilianferrari@uol.com.br.

E. Ogasawara, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, RJ – Brasil, eogasawara@ieee.org.

G. P. Guedes, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, RJ - Brasil, gustavo.guedes@cefet-rj.br.

analisar e quantificar divergências em traduções produzidas por ferramentas de TAT. Para incorporar a semântica, a métrica faz uso de um léxico psicolinguístico (afetivo), para representar as traduções candidata e referência em forma de vetores. Nesse aspecto, foi utilizado o léxico afetivo do LIWC (Linguistic Inquiry and Word Count) [14] para extrair características linguísticas e psicológicas dos textos.

Este trabalho também traz uma contribuição relevante ao comparar a métrica BRAPT ao estado da arte na avaliação de traduções (BLEU). A comparação foi efetuada a nível de sentenças (*sentence level*) dado que a BLEU também tem sido utilizada nessa tarefa [15]. Os experimentos foram realizados em um conjunto de dados que compreende 128 sentenças extraídas de dez textos jornalísticos em língua inglesa e suas respectivas traduções para o PB (Português do Brasil). Dada sua natureza, a métrica BLEU é suscetível a incompatibilizar sentenças candidatas muito parecidas com as sentenças referências. Os resultados indicam que BRAPT se mostrou mais eficaz do que a BLEU nesses casos e também em traduções que envolvem aspectos como inversões de ordem das palavras (*e.g.*, "nova namorada" e "namorada nova"), substituições de uma palavra por um sinônimo (*e.g.*, "odiar" por "detestar") e substituições de dois termos por um único termo de significado parecido (*e.g.*, "muito grande" por "enorme").

Além desta introdução, o trabalho está organizado em mais seis seções. Na Seção II, são abordados os trabalhos relacionados ao presente estudo. Na Seção III, é apresentada a fundamentação teórica, fornecendo o embasamento necessário para a compreensão da métrica proposta. A Seção IV descreve a métrica BRAPT e apresenta um algoritmo capaz de medi-la, incluindo ilustrações e fórmulas que descrevem seu funcionamento em detalhes. Os experimentos realizados são descritos na Seção V. Por fim, a Seção VI traz as considerações finais, contribuições e cenários futuros.

II. Trabalhos Relacionados

As métricas BLEU, NIST [16] e METEOR [17] são tidas como as métricas-padrão para avaliação de traduções automáticas de texto [8]. Nesse aspecto, essa seção apresenta uma breve descrição dessas métricas, assim como os problemas que podem apresentar.

A métrica BLEU pode ser utilizada para comparar sentenças de dois textos: o texto referência (*i.e.*, produzido por profissionais em traduções) e o texto candidato (*i.e.*, gerado por um tradutor automático). Essa métrica divide uma sentença em conjuntos de uma ou mais palavras dispostas de forma sequencial. Tais conjuntos são chamados de n-gramas, com o parâmetro n assumindo, em sua versão clássica, valores entre 1 e 4 [7]. O objetivo é avaliar o quão próximo o tradutor automático chega do resultado produzido pelo especialista humano, medindo a precisão modificada dos n-gramas em cada uma das sentenças. Busca-se identificar quantos conjuntos de uma única palavra (*i.e.*, unigramas) coincidem nas duas traduções. Em seguida, quantos conjuntos de duas palavras consecutivas (*i.e.*, bigramas) são coincidentes, e assim sucessivamente, até a identificação dos tetragramas que coexistem em ambas as traduções. A métrica BLEU baseia-se na suposição de que uma boa tradução tem mais n-gramas em comum com a sentença referência do que uma tradução ruim [18].

A métrica NIST é uma variação da métrica BLEU [16]. A principal diferença é que a NIST atribui pesos aos n-gramas, ou seja, quanto menos frequente for um n-grama (não importando o tamanho de n) em relação à sentença, maior é o peso dado a ele. Logo, se um n-grama correto é identificado, quanto mais raro ele for, maior é a sua importância e maior o seu peso em relação à precisão de n-gramas já que ele representa um numerador maior no cálculo da divisão efetuada para verificação da referida pontuação.

A métrica METEOR também é uma variação da BLEU. No entanto, diferente da BLEU que se concentra na precisão, a METEOR utiliza uma combinação da precisão e revocação para gerar uma média geométrica dos n-gramas em cada uma das sentenças [17]. A METEOR também considera os sinônimos por meio de palavras existentes no WordNet [19]. No entanto, foi encontrado apoio apenas para as línguas inglês, francês, alemão, espanhol, árabe e chinês [17], [20], o que inviabiliza a aplicação direta dessa métrica no contexto do português.

A métrica BLEU apresenta melhores resultados com relação ao julgamento humano de traduções [21], [22]. No entanto, por ser baseada em pareamento exato e ordenado de palavras [11], [23], essa métrica apresenta problemas para avaliar a qualidade de traduções, dado que a linguística não é uma ciência exata e que as linguagens passam por constantes modificações, mesmo em se tratando de duas traduções realizadas por especialistas. A métrica BLEU foi usada em um estudo para comparar as ferramentas Bing Tradutor e Google Tradutor, selecionando três diferentes gêneros textuais: um texto jornalístico extraído do site do Parlamento Europeu, um texto técnico referente ao manual do usuário de um notebook e um texto literário, um trecho do livro "Eat, Pray, Love", de Elizabeth Gilbert. Após analisar os resultados obtidos, percebeu-se uma sutil diferença na pontuação BLEU a favor do Google Tradutor. Essa pequena diferença, no entanto, reflete um desempenho semelhante entre os tradutores [11]. Os erros cometidos pelos tradutores também foram similares, tais como diferenças de tempos verbais, erros de conjugação, ausência ou acréscimo de artigos ou pronomes. Isso ocorre devido a uma única sentença permitir diversas traduções com semântica semelhante [11].

O estudo proposto por Rodrigues e Guedes [3] empregou 27 (das 64) categorias do LIWC para contabilizar diferenças entre traduções automáticas e traduções referências. Uma extensão desse estudo é proposta em Rodrigues et al. [12], em que os autores apresentam uma ferramenta para apresentar graficamente tais diferenças aos usuários. Ambos os estudos apenas contabilizam as diferenças da tradução candidata e da referência em termos do número de palavras que se enquadram nas categorias do LIWC. Como estudos preliminares, o objetivo consistia em mostrar que o LIWC pode ser utilizado para avaliar traduções automáticas de textos. Também é importante mencionar que a proposta dos autores não foi comparada com nenhuma outra métrica.

Além disso, esse estudo difere das métricas BLEU e NIST ao considerar a informação semântica das palavras. A nova métrica, denominada BRAPT, transcende o pareamento exato

e ordenado de palavras por ser capaz de considerar aspectos linguísticos e psicológicos contidos nas referidas sentenças com o auxílio de um léxico afetivo. É interessante ressaltar que os léxicos afetivos podem possuir informações sintáticas (*e.g.*, advérbio, verbo, pronome pessoal), psicológicas (*e.g.*, emoção positiva, emoção negativa, afeto), dentre outras.

III. Funtamentação Teórica

A. LIWC

O LIWC é uma ferramenta que contém um léxico psicolinguístico/afetivo capaz de atribuir uma ou mais categorias a palavras individuais, tornando possível contabilizar as palavras de uma sentença em categorias (e.g., processos afetivos, *function words* negações) que refletem aspectos linguísticos e psicológicos [14]. Neste estudo, utilizou-se o léxico do LIWC para representar as sentenças no espaço vetorial. Essa escolha foi devido ao fato do LIWC ser a ferramenta mais utilizada para investigar a relação entre uso de palavras e variáveis psicolinguísticas [14], [24], [25], [26]. No entanto, é importante ressaltar que existem diversos léxicos afetivos para o PB. Uma lista completa desses léxicos pode ser encontrada em Cruz et al. [27]. A Fig. 1 ilustra a representação vetorial de uma sentença s. Cada palavra $p \in s$ é contabilizada em um vetor $\vec{v}$ de acordo com a categoria x_i a qual pertence. Pode-se notar que a categoria representada pela posição x_6 obteve três palavras contabilizadas. O mesmo pode ser afirmado com relação à categoria representada pela posição x_{m-1}, por exemplo.

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	...	x_{m-1}
1	0	2	1	0	0	3	1	...	3

Fig. 1. Uma sentença s representada por um vetor $\vec{v}$, com m categorias.

B. BLEU

O objetivo da métrica BLEU é comparar as sentenças de uma tradução candidata (*i.e.*, produzida por um tradutor automático) e de uma tradução referência (*i.e.*, realizada por um especialista). Ao comparar essas sentenças, considera-se um fator chamado de precisão modificada de n-gramas, obtido por meio da divisão da quantidade de n-gramas compatíveis em ambas as sentenças pela quantidade total de n-gramas presentes na sentença candidata. Assim, ao final, considerando n = 4, obtém-se a precisão modificada de unigramas, bigramas, trigramas e tetragramas que são consideradas para o cálculo da compatibilidade BLEU. Quando a sentença candidata é menor do que a sentença referência em número de palavras, deve-se considerar uma penalidade denominada BP (Penalidade por Brevidade), pois, de acordo com essa métrica, as duas sentenças devem ser o mais semelhante possível, inclusive em tamanho. A BP deve ser contabilizada uma única vez. A Equação 1 determina a BP, em que *r* e *c* referem-se, respectivamente, à quantidade de palavras presentes nas sentenças referência e candidata.

$$BP = \begin{cases} 1, & \text{se } c \geq r \\ e^{1-\frac{r}{c}}, & \text{se } c < r \end{cases} \tag{1}$$

A BP multiplica a média geométrica das precisões modificadas, p_n, referentes a n-gramas cujos níveis variam de 1 até n e pesos positivos w_n vinculados aos seus respectivos n-gramas. As equações e explicações aqui apresentadas a respeito da métrica BLEU constam no estudo de Papineni et al. [7]. No baseline desse estudo, n está limitado a 4, visto que esse valor apresenta a melhor correlação com julgamentos humanos [7]. Os pesos $w_n = w = 1/N$ são uniformes. Considerando $n = 4$, tem-se:

$$BLEU = BP \cdot (\prod_{i=1}^{n} p_i)^{\frac{1}{n}} = BP \cdot (p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{\frac{1}{4}} \tag{2}$$

Conforme observado anteriormente, a métrica BLEU realiza uma análise baseada em pareamento exato e ordenado de palavras para avaliar a compatibilidade das sentenças. Esse tipo de análise acaba incompatibilizando traduções equivocadamente, visto que aspectos linguísticos, psicológicos e até mesmo o significado das traduções acabam sendo desconsiderados. Isso ocorre em situações como, por exemplo: inversão de ordem das palavras, substituição de palavras sinônimas (*e.g.*, "bonita" e "bela"), substituição de palavras que possuem as mesmas características linguísticas, dentre outras. Muitas vezes essas substituições não representam comprometimento significativo no entendimento da sentença que se deseja traduzir. Na Fig. 2, e, posteriormente, na Equação 3, é possível notar o impacto da BLEU ao avaliar sentenças em que houve apenas a substituição de uma palavra por um sinônimo.

n-gram	sent.	Compatibilidade BLEU (Referência vs Candidata)												p_n
1-gram	Ref.	Ela	vai	odiar	meu	novo	carro							5/6
	Cand.	Ela	vai	detestar	meu	novo	carro							
2-gram	Ref.	Ela vai		vai odiar		odiar meu		meu novo		novo carro				3/5
	Cand.	Ela vai		vai detestar		detestar meu		meu novo		novo carro				
3-gram	Ref.	Ela vai odiar			vai odiar meu			odiar meu novo			meu novo carro			1/4
	Cand.	Ela vai detestar			vai detestar meu			detestar meu novo			meu novo carro			
4-gram	Ref.	Ela vai odiar meu				vai odiar meu novo				odiar meu novo carro				0/3
	Cand.	Ela vai detestar meu				vai detestar meu novo				detestar meu novo carro				

Fig. 2. Exemplo de cálculo da métrica BLEU.

$$BLEU = BP \cdot (p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{\frac{1}{4}}$$

$$BLEU = 1 \cdot \left(\frac{5}{6} \cdot \frac{3}{5} \cdot \frac{1}{4} \cdot \frac{0}{3}\right)^{\frac{1}{4}} = 0 \tag{3}$$

$$BLEU = 1 \cdot (0,83 \cdot 0,6 \cdot 0,25 \cdot 0)^{0,25} = 0$$

Na ilustração da Fig. 2, todas as incompatibilidades estão destacadas em amarelo e notadamente causam impacto negativo na precisão modificada de *n-gramas* (representada pela coluna p_n). Torna-se importante observar que, além da troca de palavras não representar um comprometimento significativo da tradução, basta haver o comprometimento total de uma única sequência de *n-gramas*, como foi o caso dos tetragramas, por exemplo, para que a métrica BLEU apresente total incompatibilidade (BLEU = 0).

IV. A MÉTRICA BRAPT

Em linhas gerais, a métrica BRAPT (*Bilingual Rating of Psycholinguistic Perspectives in Translations*), objeto principal deste trabalho, consiste em representar as sentenças referência e candidata na forma de vetores, em que cada uma de suas posições representa uma categoria que pode refletir aspectos linguísticos ou psicológicos do léxico do LIWC em PB, que contém 64 categorias. É importante ressaltar que, nesse estudo, todas as 64 categorias foram utilizadas. Dessa maneira, os referidos vetores possuem, em cada uma de suas posições, um valor inteiro representando a quantidade de palavras contabilizadas por categoria. Como exemplo ilustrador, podemos destacar a palavra "odiar", que se enquadra nas categorias: verbo, processo afetivo, emoção negativa e raiva. Cada vetor $\vec{v}$ contém $m+1$ posições correspondentes à quantidade de categorias existentes no léxico do LIWC em PB (*i.e.*, $m = 64$ categorias), com a adição de uma categoria extra para contabilizar as palavras não existentes no referido léxico. Assim, cada sentença s é representada por um vetor $\vec{v}$ de 65 posições. Uma vez que as duas sentenças se encontram representadas na forma de vetores de categorias, o próximo passo é verificar a similaridade entre os mesmos, ou seja, a similaridade entre as duas representações de sentenças.

Para calcular a similaridade (*i.e.*, compatibilidade) entre duas sentenças (*i.e.*, dois vetores), foi utilizada a similaridade do cosseno. A similaridade do cosseno é uma das medidas de similaridade mais comuns e mais estudas [28]. Sua utilização é preferida com relação a outras métricas (e.g., Jaccard, Euclidiana) dado que normaliza o comprimento do texto durante a comparação entre textos [28], [29]. Essa medida é conhecida na literatura e há diversos estudos que a utilizam em mineração de textos e PLN para comparar textos ou sentenças e verificar sua similaridade. Dentre esses estudos, é possível destacar os trabalhos de Santos et al. [16] e de Di Thommazo et al. [30]. A similaridade do cosseno é calculada a partir do produto interno entre dois vetores que representam textos ou sentenças, como é o caso do presente estudo. Essa similaridade representa o ângulo entre os dois vetores [31]. Logo, os valores possíveis para a similaridade variam entre 0 (*i.e.*, ausência completa de similaridade) e 1 (*i.e.*, similaridade total). O cômputo do BRAPT é fornecido pelo algoritmo calcBRAPT, apresentado no Algoritmo 1.

Algoritmo 1 calcBRAPT(s_{ref}, s_{cand}, lex)

Input:

- s_{ref} = Sentença referência
- s_{cand} = Sentença candidata
- lex = Léxico afetivo

Output: S, a similaridade entre s_{ref} e s_{cand}

1: $\mathrm{s} \leftarrow \emptyset$
2: $\vec{v}_{ref} \leftarrow \mathrm{termosPorCat}(s_{ref}, lex)$
3: $\vec{v}_{cand} \leftarrow \mathrm{termosPorCat}(s_{cand}, lex)$
4: $\mathrm{S} \leftarrow \mathrm{sim}(\vec{v}_{ref}, \vec{v}_{cand})$
5: **return** S

Nos passos 2 e 3 do calcBRAPT é possível observar o cômputo dos valores contidos nas posições dos vetores $\vec{v}_{ref}$ e $\vec{v}_{cand}$, que representam as sentenças referência e candidata, respectivamente. A função `termosPorCat` recebe como argumento, em momentos distintos, duas sentenças (*i.e.*, s_{ref} ou s_{cand}) e retorna um vetor de $m+1$ posições para cada sentença. Dado que léxicos como o LIWC podem ter uma palavra relacionada a uma ou mais categorias, o vetor $\vec{v}_{ref}$ é representado de maneira que cada palavra contida em s_{ref} possa ser contabilizada em uma ou mais categorias x_i. O mesmo ocorre com $\vec{v}_{cand}$ e s_{cand}. Os passos 4 e 5 representam a verificação e o retorno da similaridade entre os dois vetores.

A Fig. 3 ilustra o cômputo da similaridade entre dois vetores (*i.e.*, duas sentenças). Os vetores x e y poderiam ser referentes, respectivamente, às sentenças referência e candidata. O algoritmo calcBRAPT opera em função do número de categorias do léxico utilizado, que nesse caso hipotético é cinco (*i.e.*, $m+1 = 6$ posições).

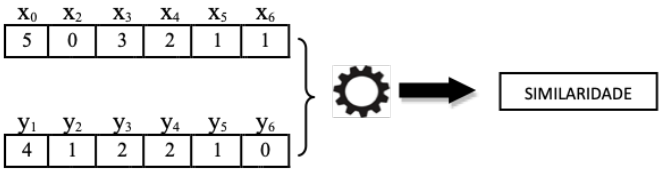


Fig. 3. Similaridade entre dois vetores para $m = 5$.

A similaridade do cosseno verificada entre os vetores x e y é representada por S e pode ser obtida pela equação 4. A referida equação e o desenvolvimento dos cálculos resultantes de sua aplicação são efetuados em seguida. O valor da similaridade S resulta em $0,96$.

$$S = \frac{x_i \cdot y_i}{\sqrt{\sum_1^n x_i^2} \times \sqrt{\sum_1^n y_i^2}} = \frac{31}{6,33 \times 5,10} = 0,96 \qquad (4)$$

V. EXPERIMENTOS

Para avaliar a métrica proposta, foram selecionados dez textos jornalísticos em inglês, que totalizam 128 sentenças. Esses textos foram traduzidos para o PB por profissionais especializados (especialistas), produzindo o que se conhece por traduções referência e, em seguida, traduzidos por três ferramentas de TAT conhecidas na *web*: GT (Google Tradutor), BI (Bing Tradutor) e WL (WorldLingo Tradutor), produzindo o que se conhece por traduções candidatas. As traduções referência e candidata foram submetidas às métricas BRAPT e BLEU.

A escolha dos especialistas atendeu ao seguinte critério: deveriam ser professores com mais de 10 anos de atividades profissionais em traduções de inglês/português. Foi solicitado aos especialistas que retornassem as traduções em no máximo 30 dias. Dado que se tratavam de pessoas experientes, não houve recomendações sobre o material que poderiam consultar ou ferramentas que poderiam utilizar. Além disso, como os especialistas se encontravam em diferentes regiões, não houve determinações sobre o local em que fariam as traduções. Por fim, o envio dos textos originais foi feito por e-mail e, da mesma maneira, as traduções foram retornadas por e-mail.

A avaliação experimental foi organizada em três partes. Na primeira, indicada na seção V.A, foi realizada uma análise

comparativa das traduções feitas pelo GT, BI e WL frente a uma tradução referência. Posteriormente é feita (seção V.B) uma análise comparativa do índice de compatibilidade medidos pela BLEU e BRAPT frente à visão de diferentes especialistas. Finalmente, na seção V.C, é feita uma análise detalhada por sentença de modo a facilitar a compreensão da similaridade computada pela BLEU e pela BRAPT.

A. *Comparação de Traduções*

Essa seção apresenta os resultados de uma análise texto a texto (*i.e.*, considerando a compatibilidade média entre as sentenças de cada texto) utilizando as métricas BLEU e BRAPT. Esses resultados objetivam avaliar o desempenho das ferramentas BI, GT e WL e são exibidos, respectivamente, na Fig. 4 e Fig. 5. A partir do uso das duas métricas (BLEU e BRAPT), observa-se que as ferramentas GT e BI apresentam um desempenho parelho, ou seja, com traduções candidatas mais semelhantes às traduções referência, e superior ao desempenho da ferramenta WL em cada texto e na média geral. Em alguns casos, é possível notar que a proporção da compatibilidade BRAPT é diferente da compatibilidade BLEU, por exemplo, nos textos 4, 7 e 9 (*Tex4*, *Tex7* e *Tex9*, respectivamente). Nos textos 4 e 7, a compatibilidade BLEU apresenta aproximadamente 9% de distanciamento entre o BI e GT (Figura 4), o que não acontece na compatibilidade BRAPT, que apresenta esse distanciamento bastante reduzido (Figura 5). No caso do texto 9, o WL se distancia bastante do BI e GT na compatibilidade BLEU, o que não ocorre na compatibilidade BRAPT. Isso ocorre porque a BRAPT envolve a semântica das palavras, o que tende a reduzir a diferença nas traduções com palavras que possuem conteúdo semântico semelhante. Independente da redução das diferenças, é importante ressaltar que os gráficos se comportam de maneira semelhante ao considerar a comparação entre as duas métricas, ou seja, sempre que a ferramenta se comportou melhor na compatibilidade BLEU, se comportou melhor na compatibilidade BRAPT. Isso pode ser observado, por exemplo, com o *Tex2*, que apresenta a melhor compatibilidade sendo o BI, em seguida o GT e por fim o WL.

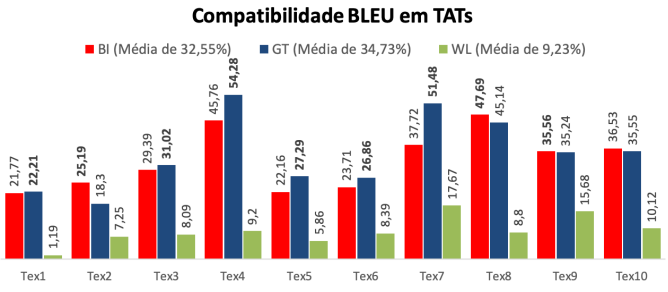


Fig. 4. Compatibilidade média dos textos de acordo com a BLEU.

Ao analisar as ferramentas de TAT com base nas 128 sentenças, a métrica BRAPT indica que as ferramentas GT e BI apresentam traduções candidatas bem mais similares às traduções referências do que a ferramenta WL, que apresentou um melhor desempenho em apenas 6% das sentenças. Os resultados podem ser observados na Fig. 6, que sinaliza uma

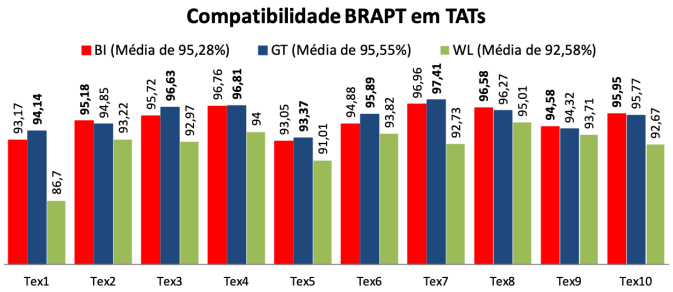


Fig. 5. Compatibilidade média dos textos de acordo com a BRAPT.

vantagem muito discreta para o GT. Pode-se destacar também que, segundo a métrica BRAPT, o GT apresenta melhor desempenho em 42% das traduções enquanto o BI apresenta 41%. É interessante ressaltar que houve empate entre o GT e o BI em 11% das traduções.

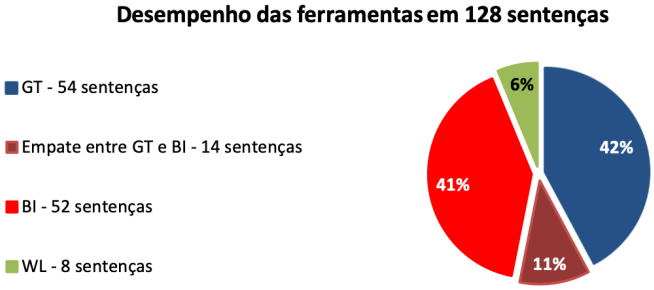


Fig. 6. Desempenho das ferramentas de TAT de acordo com a BRAPT.

Os dados apresentados são consistentes com o que já existe na literatura sobre a análise das ferramentas BI e GT, haja vista os resultados apresentados no estudo de Melo et al. [11]. Tais textos também foram analisados pelos especialistas, que confirmaram que as traduções da ferramenta WL foram inferiores às traduções das outras duas ferramentas. Segundo as impressões dos mesmos, as ferramentas BI e GT apresentam desempenho parelho e produzem traduções bem próximas às traduções referência em grande parte das sentenças, chegando a apresentar traduções iguais em alguns casos.

B. *Análise do Índice de Compatibilidade*

Os resultados apresentados pela métrica BLEU indicaram percentuais relativamente inferiores à métrica BRAPT. Visando analisar a valoração da compatibilidade, os resultados produzidos pelas métricas BRAPT e BLEU foram comparados com as impressões dos especialistas, que utilizaram uma escala (*rating scale*) de 0 a 10 para avaliarem cada sentença no texto 6, conforme apresentado na Fig. 7.

Além da disparidade entre os valores BLEU e BRAPT, pode-se notar que a BRAPT apresenta números mais próximos das avaliações dos especialistas (barras laranjas) em todas as sentenças. Vale destacar a incompatibilidade total de algumas sentenças segundo a BLEU (s4, s5, s7 e s16). Dentre essas quatro citadas, a menor (s7) e a maior (s16) sentença foram utilizadas para uma análise minuciosa, exibida na Tabela I.

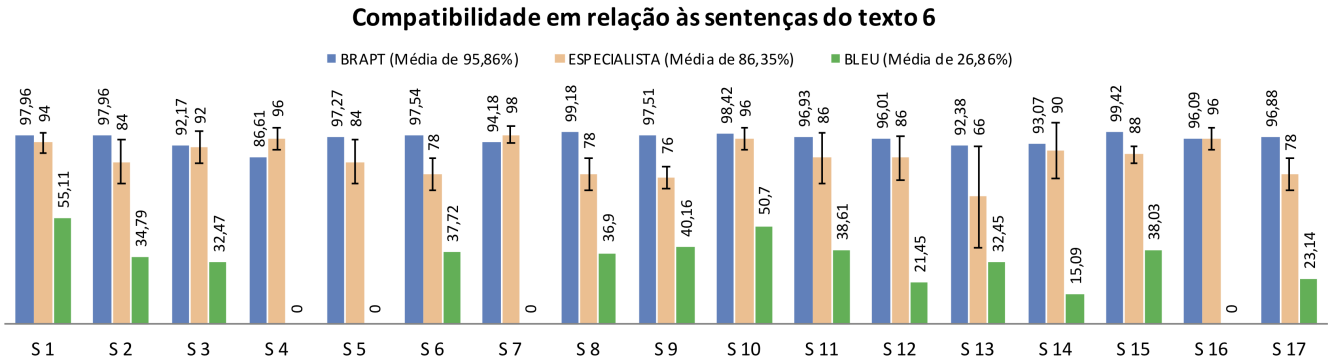


Fig. 7. Compatibilidade verificada em sentenças traduzidas pelo GT.

TABELA I
SENTENCAS 7 E 16 DO TEXTO 6: BLEU VS BRAPT

Referência: Mas não o são. **Candidata (GT):** Mas **eles** não são.		
Especialistas = 98%	**BRAPT = 94,2%**	**BLEU = %**

Referência: Felizmente seu estudo **da** relatividade a preparou para o choque **que terá ao ver** sua irmã gêmea agora **com** 71 anos. **Candidata (GT):** Felizmente seu estudo **sobre a** relatividade preparou**-a** para o choque **quando vê** sua irmã gêmea **que** agora **tem** 71 anos.		
Especialistas = 96%	**BRAPT = 96,1%**	**BLEU = 0%**

TABELA II
COMPATIBILIDADE: BLEU VS BRAPT

Referência 1: Ela vai **odiar** meu novo carro **Candidata 1:** Ela vai **detestar** meu novo carro		
Especialistas = 100%	**BRAPT = 93,8%**	**BLEU = 0%**

Referência 2: Minha **nova namorada** é muito alta **Candidata 2:** Minha **namorada nova** é muito alta		
Especialistas = 94%	**BRAPT = 100%**	**BLEU = 0%**

Referência 3: A casa é **muito grande** **Candidata 3:** A casa é **enorme**		
Especialistas = 98%	**BRAPT = 79.21%**	**BLEU = 0%**

As sentenças acima explicitam o rigor da métrica BLEU ao descartar (*i.e.*, compatibilidade = 0%) sentenças candidatas com significados muito próximos da sentença referência. Na métrica BRAPT, bem como na visão dos especialistas, essas variações caracterizam apenas uma redução no percentual de compatibilidade ao invés de descartar as referidas sentenças.

C. Análise Detalhada por Sentença

Também foram realizados experimentos com sentenças que foram avaliadas de forma ineficaz pela métrica BLEU. Esses experimentos visaram analisar aspectos como a substituição de uma palavra por um sinônimo, a inversão de ordem das palavras e a avaliação de candidatas menores com substituição de dois termos por um termo de significado parecido (*e.g.*, "muito grande" por "enorme"), novamente a métrica BRAPT apresentou percentuais mais condizentes com o significado das sentenças referência e candidata avaliadas. Esses aspectos podem ser observados em maiores detalhes, na Tabela II.

Os exemplos acima ilustram a dificuldade apresentada pela métrica BLEU para avaliar aspectos que transcendem o pareamento exato e ordenado de palavras. É possível notar que nos três exemplos a métrica BLEU apresentou incompatibilidade total (0%).

Em relação à métrica BRAPT, no primeiro exemplo, as palavras "odiar" e "detestar" foram contabilizadas nas categorias *verb*, *affect*, *negemo* e *anger*. A única divergência foi a contabilização da palavra "detestar" na categoria *humans*. A palavra "odiar" não se enquadra nessa categoria e por isso houve uma perda de compatibilidade. Como todas as demais palavras são iguais, não houve outras divergências. No segundo exemplo, como essa inversão de ordem não interfere na compatibilidade BRAPT a compatibilidade foi de 100%. No terceiro exemplo, as palavras "grande" e "enorme" se enquadram em categorias parecidas. No entanto, a palavra "muito", ausente na candidata, se enquadra em diversas categorias (*e.g.*, *adverb.*, *quant*, *discrep*), causando apenas a diminuição da compatibilidade (*i.e.*, $79,21\%$).

VI. CONCLUSÃO

O presente trabalho propõe a métrica BRAPT para avaliar similaridade de traduções de textos. O trabalho apresenta uma análise comparando o BRAPT ao estado da arte (métrica BLEU). A métrica BLEU, por ser baseada em pareamento exato e ordenado de palavras, mostrou-se limitada. Essa métrica desconsidera aspectos importantes das traduções, como a inversão de ordem de palavras, utilização de sinônimos, dentre outros. Tais limitações são compensadas no BRAPT pela semântica das palavras.

Na avaliação experimental, por meio de prova de conceito, a métrica BRAPT mostrou-se mais adequada do que a métrica BLEU, tanto pela particular capacidade de considerar os aspectos linguísticos e psicológicos das traduções, quanto pela avaliação mais flexível e menos suscetível a invalidar traduções candidatas semelhantes às traduções referências. Tal resultado foi confirmado por meio de análises feitas por especialistas. Em trabalhos futuros, pode-se utilizar a métrica BRAPT com outras versões do LIWC ou com outros léxicos afetivos similares, além de outras técnicas para explorar outras medidas de similaridade de textos. Além disso, objetivamos

utilizar a métrica proposta para avaliar traduções automáticas em outros idiomas. Também pretende-se comparar a métrica BRAPT com outras métricas presentes na literatura.

Diante do cenário apresentado, a métrica proposta pode ser utilizada para contornar o problema de traduções que podem ter conteúdo semântico semelhante, mas que não são bem avaliadas por métricas de avaliação de traduções automáticas de texto. Nesse aspecto, pode ser adotada, por exemplo, como métrica para avaliar sistemas de TAT e em sistemas de aprendizagem *online* para o treinamento de especialistas tradutores [32]. Como limitação dessa proposta, entende-se que está condicionada à existência de um léxico do LIWC para a língua correspondente. No entanto, novas versões têm surgido, dentre elas, pode-se destacar as línguas: chinês [33], japonês [34], holandês [35], dentre outras.

Por fim, entende-se que este trabalho preenche uma lacuna no tocante à avaliação de TATs sob aspectos linguísticos e psicológicos ao considerar a semântica das palavras no processo de comparação de sentenças traduzidas. As novas perspectivas apresentadas representam uma contribuição relevante e apontam para uma nova direção a ser seguida em outras pesquisas nessa área.

AGRADECIMENTOS

Os autores agradecem ao CNPq, CAPES (código de financiamento 001) e FAPERJ por parcialmente auxiliar nessa pesquisa.

REFERÊNCIAS

[1] E. Liddy, "Natural Language Processing," in *Encyclopedia of Library and Information Science*, 2nd ed., M. Drake, Ed. New York: CRC Press, 2001.

[2] A. Gelbukh, G. Sidorov, S.-Y. Han, and E. Hernández-Rubio, "Automatic Enrichment of a Very Large Dictionary of Word Combinations on the Basis of Dependency Formalism," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, vol. 2972, 2004, pp. 430–437.

[3] R. Rodrigues, R. Gomes, K. Rodrigues, and G. Guedes, "TATMaster: Psycholinguistic Divergences in Automatically Translated Texts," in *WebMedia 2017 - Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, 2017, pp. 205–208.

[4] R. Rocha, "Adaptive Technology Applied to Natural Language Processing," *IEEE Latin America Transactions*, vol. 5, no. 7, pp. 544–551, 2007.

[5] O. Başkaya, E. Yildiz, D. Tunaoğlu, M. Tolga Eren, and A. Seza Doğruöz, "Integrating Meaning into Quality Evaluation of Machine Translation," in *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, vol. 1, 2017, pp. 210–219.

[6] E. H. Hovy, "Toward Finely Differentiated Evaluation Metrics for Machine Translation," in *Proceedings of the Eagles Workshop on Standards and Evaluation*, Pisa, Italy, 1999.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[8] X. Zeng, L. Chao, D. Wong, I. Trancoso, and L. Tian, "Toward Better Chinese Word Segmentation for SMT via Bilingual Constraints," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, vol. 1, 2014, pp. 1360–1369.

[9] D. Cer, C. Manning, and D. Jurafsky, "The Best Lexical Metric for Phrase-Based Statistical MT System Optimization," in *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 2010, pp. 555–563.

[10] K. Papineni, S. Roukos, T. Ward, J. Henderson, and F. Reeder, "Corpus-Based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results," in *Proceedings of the Second International Conference on Human Language Technology Research*, ser. HLT '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 132–137.

[11] F. R. d. Melo, H. C. d. O. Matos, and E. R. B. Dias, "Aplicação da Métrica BLEU para Avaliação Comparativa dos Tradutores Automáticos Bing Tradutor e Google Tradutor," *Revista e-scrita: Revista do Curso de Letras da UNIABEU*, vol. 5, no. 3, pp. 33–45, Jan. 2015.

[12] R. Rodrigues and G. Paiva Guedes, "TATModel: Em Direção a um Novo Modelo para Avaliação de Traduções Automáticas de Texto," in *Proceedings of the 5th Symposium on Knowledge Discovery, Mining and Learning (KDMile)*, Uberlândia, MG, Brazil, 2017, pp. 161–164.

[13] K. Wołk, W. Glinkowski, and A. Żukowska, "Enhancing the Assessment of (Polish) Translation in PROMIS Using Statistical, Semantic, and Neural Network Metrics," *Advances in Intelligent Systems and Computing*, vol. 746, pp. 351–366, 2018.

[14] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count*, 2nd ed. Psychology Press, Feb. 2001.

[15] M. Stanojevic and K. Sima'an, "BEER: BEtter Evaluation as Ranking," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 414–419.

[16] J. Santos, I. Anastácio, and B. Martins, "Named Entity Disambiguation over Texts Written in the Portuguese or Spanish Languages," *IEEE Latin America Transactions*, vol. 13, no. 3, pp. 856–862, 2015.

[17] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation*, ser. StatMT '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 228–231.

[18] A. Finch, Y. Akiba, and E. Sumita, "How Does Automatic Machine Translation Evaluation Correlate with Human Scoring as the Number of Reference Translations Increases?" in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004.

[19] G. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[20] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Comparative Study Between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, 2015.

[21] V. López-Ludeña, R. San-Segundo, J. Montero, R. Córdoba, J. Ferreiros, and J. Pardo, "Automatic Categorization for Improving Spanish into Spanish Sign Language Machine Translation," *Computer Speech and Language*, vol. 26, no. 3, pp. 149–167, 2012.

[22] Z. M. Almahasees, "Assessing the Translation of Google and Microsoft Bing in Translating Political Texts from Arabic into English," *International Journal of Languages, Literature and Linguistics*, vol. 3, no. 1, pp. 1–4, Mar. 2017.

[23] J. C. Gomez, T. Tommasi, S. Zoghbi, and M. F. Moens, "What Would They Say? Predicting User's Comments in Pinterest," *IEEE Latin America Transactions*, vol. 14, no. 4, pp. 2013–2019, Apr. 2016.

[24] J. Mahmud, "Why do You Write This? Prediction of Influencers from Word Use," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[25] M. R. Mehl and A. J. Gill, "Computerized Content Analysis," in *Advanced methods for behavioral research on the internet*, Washington DC, 2010.

[26] A. J. Gill, A. Vasalou, C. Papoutsi, and A. Joinson, "Privacy dictionary: A Linguistic Taxonomy of Privacy for Content Analysis," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 3227–3236, 2011.

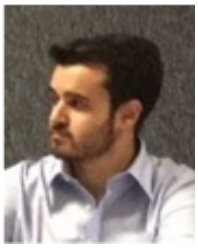
[27] P. P. Cruz, R. Rodrigues, K. Belloze, and G. P. Guedes, "Uma Revisão Sistemática sobre Léxicos Afetivos para o Português do Brasil," in *XXIII Conferência Internacional sobre Informática na Educação (TISE2017)*, Fortaleza, Brazil, 2017.

[28] A. Park, A. Hartzler, J. Huh, D. McDonald, and W. Pratt, "Homophily of Vocabulary Usage: Beneficial Effects of Vocabulary Similarity on Online Health Communities Participation," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2015, pp. 1024–1033, 2015.

[29] J. Prominski, P. Shah, and R. Alvarado, "Editor Matching for Academic Journals Through Rich Semantic Network Development," in *2018 Systems and Information Engineering Design Symposium, SIEDS 2018*, 2018, pp. 287–292.

[30] A. Di Thommazo, G. Malimpensa, T. De Oliveira, G. Olivatto, and S. Fabbri, "Requirements Traceability Matrix: Automatic Generation and Visualization," in *Proceedings - 2012 Brazilian Symposium on Software Engineering, SBES 2012*, 2012, pp. 101–110.
[31] M. G. de Oliveira, E. Oliveira, and R. Z. Marchesi, "Um QAsystem para Interação de Alunos em Avaliações Somativas a Distância," *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, vol. 1, no. 1, Nov. 2009.
[32] C. Rösener, "A Linguistic Intelligent System for Technology Enhanced Learning in Vocational Training - The ILLU project," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5794 LNCS, pp. 800–805, 2009.
[33] C.-L. Huang, C. K. Chung, N. Hui, Y.-C. Lin, Y.-T. Seih, B. C. P. Lam, W.-C. Chen, M. H. Bond, and J. W. Pennebaker, "The Development of the Chinese Linguistic Inquiry and Word Count Dictionary," *Chinese Journal of Psychology*, vol. 54, no. 2, pp. 185–201, 2012.
[34] D. Shibata, S. Wakamiya, A. Kinoshita, and E. Aramaki, "Detecting Japanese Patients with Alzheimer's Disease based on Word Category Frequencies," in *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 78–85.
[35] P. Boot, H. Zijlstra, and R. Geenen, "The Dutch Translation of the Linguistic Inquiry and Word Count (LIWC) 2007 Dictionary," *Dutch Journal of Applied Linguistics*, vol. 6, no. 1, pp. 65–76, 2017.

Rafael Guimarães Rodrigues é graduado em Análise e Desenvolvimento de Sistemas pela Faculdade de Filosofia Santa Dorotéia (FFSD), Nova Friburgo, Rio de Janeiro, Brasil, em 2001. Possui Pós-graduação em Administração de Banco de Dados pela Universidade Estácio de Sá (UNESA), Rio de Janeiro, Brasil, em 2008. Obteve o título de Mestre em Ciência da Computação pelo Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, Rio de Janeiro, Brasil, em 2018. Foi professor da Faculdade de Filosofia Santa Dorotéia entre 2006 e 2013. Atualmente é professor do CEFET/RJ, campus Nova Friburgo, Rio de Janeiro, Brasil, desde setembro de 2012. Suas pesquisas concentram-se na área de computação afetiva.



Kaio Tavares Rodrigues é graduando em Relações Internacionais pelo Instituto de Estudos Estratégicos da Universidade Federal Fluminense (INEST-UFF), com previsão de término para 2019, Niterói, Rio de Janeiro, Brasil. Atua, desde 2015, como pesquisador de IC vinculado ao Laboratório de Estudos sobre Política Externa Brasileira (LEPEB/INEST). É também tradutor formado pelo Curso de Formação de Tradutores Daniel Brilhante de Brito (DBB), Rio de Janeiro, Brasil, 2014. Trabalha, desde então, com traduções literárias de inglês e castelhano.

Rodrigo Reis Gomes é graduado em Bacharelado em Informática pela Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil, em 1999. Obteve os títulos de Mestre e Doutor em Modelagem Computacional pela Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, RJ, Brasil, em 2003 e 2012, respectivamente. Foi professor da Universidade Estácio de Sá (UNESA) entre 2002 e 2008. Desde então, é professor do CEFET/RJ, campus Nova Friburgo, RJ, Brasil.

Lilian Ferrari é graduada em Psicologia pela Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brasil, em 1980. Obteve os títulos de Mestre e Doutora em Linguística pela Universidade Federal do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brasil, em 1985 e 1994, respectivamente. É Professora Titular do Departamento de Linguística da UFRJ, Pesquisadora do CNPq e líder do Laboratório de Linguística Cognitiva (LINC/UFRJ).

Eduardo Ogasawara possui bacharelado em Ciência da Computação pela UFRJ em 1997. Obteve os títulos de Mestre e Doutor em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil, em 2000 e 2011. Atualmente é professor do CEFET/RJ, desde de 2010 e é Pesquisador do CNPq. Tem sólida formação em Banco de Dados e sua pesquisa concentra-se em Ciência de Dados.

Gustavo Paiva Guedes possui bacharelado em Ciência da Computação pelo Centro Universitário Plinio Leite, Niterói, Rio de Janeiro, Brasil, em 2004. Além disso, possui bacharelado em Letras (Português/Alemão) pela UFRJ, em 2005. Possui pós-graduação em Gerência de Tecnologia em Computação pela UFF, em 2006. Obteve seu mestrado em Linguística pela UFRJ, em 2008 e o doutorado em Engenharia de Sistemas e Computação também pela UFRJ, em 2015. Atualmente é professor do CEFET/RJ, desde outubro de 2010. Suas pesquisas concentram-se na área de computação afetiva.