CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ASPECTOS SEMÂNTICOS EM TRADUÇÕES AUTOMÁTICAS DE TEXTOS

Rafael Guimarães Rodrigues

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador:
Gustavo Paiva Guedes e Silva

Rio de Janeiro,

Maio 2018

Aspectos Semânticos em Traduções Automáticas de Textos

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Rafael Guimarães Rodrigues

Aprovada por:

Presidente, Prof. Gustavo Paiva Guedes e Silva, D.Sc. (orientador)

Prof. Eduardo Soares Ogasawara, D.Sc.

Prof. Eduardo Bezerra da Silva, D.Sc.

Profª. Lilian Vieira Ferrari, D.Sc. (Universidade Federal do Rio de Janeiro)

Rio de Janeiro,

Maio 2018

CEFET/RJ – Sistema de Bibliotecas / Biblioteca Campus Nova Friburgo

R696a Rodrigues, Rafael Guimarães
Aspectos Semânticos em traduções automáticas de textos / Rafael Guimarães Rodrigues. — 2018.
xiv, 78f. : il.color. , grafs. , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2018.
Bibliografia : f. 69-73
Orientador : Gustavo Paiva Guedes e Silva, D.Sc.

1. Psicolinguística. 2. Semântica. 3. Tradução automática de texto - Análise. 4. Ciência da Computação. I. Silva, Gustavo Paiva Guedes e (Orient.). II. Título.

CDD 401.9

Elaborada pela bibliotecária Cristina Rodrigues Alves CRB-7/5932

A meus pais e minha irmã por tudo.

A sabedoria é um paradoxo. O homem que mais sabe é aquele que mais reconhece a vastidão da sua ignorância.

Friedrich Nietzsche

Agradecimentos

À Deus, por me acompanhar e me guiar em cada passo dessa caminhada;
À minha família, responsável por tudo de melhor que há em mim;
À Gustavo Paiva Guedes, por tornar essa dissertação possível, por sua orientação, dedicação, amizade, generosidade, paciência e enorme contribuição para minha evolução profissional;
Aos professores Kele Belloze e Joel dos Santos pelo profissionalismo, boa vontade e excelente trabalho desenvolvido nos seminários que tanto me ajudaram ao longo do curso;
Ao professor Eduardo Ogasawara pelas valiosas contribuições em minha qualificação e ao longo do curso, pelo incentivo, presteza e pelo trabalho de excelência na coordenação do PPCIC;
Aos professores Eduardo Bezerra e Lilian Ferrari, por participarem desta banca e pelas valiosas contribuições em minha qualificação.
Aos demais professores do PPCIC pelos conselhos e ensinamentos ao longo do curso;
À todos os meus colegas de curso pelo essencial apoio nos momentos mais difíceis;
À Mário Roberto Ferreira de Lima (*in memorian*), meu grande exemplo na profissão, por ter me iniciado na vida acadêmica e por seus valiosos conselhos que levarei para toda a vida;
À irmã Celma Calvão, Rozária Heller e todos os colegas e funcionários da saudosa Faculdade Santa Dorotéia, pela minha formação acadêmica e minha carreira como docente;
À Rodrigo Reis Gomes, amigo de todas as horas, pela boa vontade em ajudar com as formalizações e pelo incentivo de sempre;
À Gabriel Cornélio de Moura, Thiago Delgado Pinto, Paulo Henrique Werly, Dalmo Stutz, Isabel Spitz Lavandier e Dacy Câmara Lobosco pelas longas conversas sobre essa caminhada;
À Kaio Tavares Rodrigues pela amizade e inestimável contribuição profissional;
À todos os meus amigos mais próximos pela torcida, pelas conversas, pelo apoio, pela compreensão de minha ausência e pelas orações;
Às pessoas que trabalham no CEFET e que de alguma forma me ajudaram nessa jornada;
À Léa Freitas, Daniel Ribeiro, Jorge Inaba, Gisele Breder e Ana Paula Huback pelas inestimáveis contribuições;
Por fim, a todos os docentes que contribuíram para a minha formação desde o início;

agradeço.

RESUMO

Aspectos Semânticos em Traduções Automáticas de Textos

Rafael Guimarães Rodrigues

Orientador:
Gustavo Paiva Guedes e Silva

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do título de mestre.

As traduções automáticas de texto surgiram nos anos 50, motivadas por questões militares. Atualmente esse tipo de tradução faz parte do nosso cotidiano e representa uma importante ferramenta para a comunicação no mundo globalizado, especialmente com a utilização de ferramentas de tradução automática de textos disponíveis em ambiente *web*. No entanto, apesar de tratar-se de uma área com mais de 60 anos de estudos, ainda há diversos desafios a serem superados, o que faz com que esse tipo de processo continue dependente de revisão humana. Existem, atualmente, diversas métricas para avaliar traduções automáticas de textos, dentre as quais, a métrica BLEU apresenta-se como o estado da arte. Essa métrica avalia a qualidade das traduções com base no pareamento exato e ordenado de palavras, sem considerar, no entanto, a semântica (*e.g.,* aspectos linguísticos e psicológicos) das sentenças avaliadas. Nesse cenário, o principal objetivo deste trabalho é propor uma nova métrica capaz de adicionar semântica às avaliações desse tipo de tradução. Como objetivo secundário, esse trabalho também contribui com dois algoritmos para auxiliar na identificação e quantificação de aspectos psicolinguísticos em traduções do inglês para o português do Brasil. Para alcançar os objetivos propostos, este trabalho utiliza um léxico afetivo presente em uma ferramenta denominada LIWC (*Linguistic Inquiry and Word Count*). Esse léxico é capaz de contabilizar palavras em categorias que representam aspectos psicológicos e linguísticos. Durante os experimentos foram utilizados dez textos traduzidos por dois especialistas humanos e por três dessas ferramentas já citadas. Os referidos textos foram utilizados para estabelecer uma comparação entre a métrica proposta e o estado da arte. Os testes também objetivaram avaliar possíveis problemas produzidos por ferramentas utilizadas para realizar esse tipo de tradução. Os resultados foram considerados promissores e indicam que esse estudo pode contribuir com novos trabalhos direcionados ao desenvolvimento de métricas para avaliação de traduções automáticas de textos e talvez até mesmo para trabalhos direcionados ao desenvolvimento de ferramentas que produzam esse tipo de tradução.

Palavras-chave:
Traduções automáticas de textos;Semântica; Léxico afetivo.

Rio de Janeiro,

Maio 2018

ABSTRACT

Semantic Aspects in Automatic Texts Translations

Rafael Guimarães Rodrigues

Advisor:
Gustavo Paiva Guedes e Silva

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Automatic text translations appeared in the fifties, motivated by military questions. Nowadays this type of translation is part of our everyday life and represents an important tool for communication in the globalized world, especially with the use of automatic translation tools available in texts. However, although it is an area with more than 60 years of study, there are still several challenges to be overcome, which makes this type of process remain dependent on human review. There are currently several metrics to evaluate automatic translations of texts, among which, the BLEU metric is presented as the state of the art. This metric evaluates the quality of the translations based on the exact and ordered pairing of words, without considering, however, semantic (linguistic and psychological aspects) of the sentences. In this scenario, the primary objective of this work is to propose a new metric capable of adding semantics to the evaluations of this type of translation. As a secondary objective, this work also contributes with two algorithms to help to identify and to quantify psycholinguistic aspects of translations from English to Brazilian Portuguese. To reach the proposed objectives, this work uses an affective lexicon present in a tool named LIWC (Linguistic Inquiry and Word Count). This lexicon is capable of counting words in categories that represent psychological and linguistic aspects. During the experiments, ten translated texts by two human experts and three of these tools were used. These texts were used to establish a comparison between the proposed metric and BLEU. The tests also aimed to evaluate possible problems produced by tools used to perform this type of translation. The results were considered promising and indicate that this study can contribute with new works directed to the development of metrics for the evaluation of automatic translations of texts and perhaps even for works directed to the development of tools that produce this type of translation.

Key words:
Automatic translation of texts; Semantic; Affective lexicon.

Rio de Janeiro,

11 de junho de 2018

Rodrigues, Rafael Guimarães.

Aspectos Semânticos em Traduções Automáticas de Textos / Rafael Guimarães Rodrigues – 2018.
x, 73 f; enc.

Dissertação (Mestrado), Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2018.
Bibliografia: f, 69–73

1. Divergências psicolinguísticas 2. Traduções automáticas de textos I. Título

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

BLEU *Bilingual Evaluation Understudy* . . . 2, 3, 4, 5, 7, 11, 12, 13, 15, 16, 17, 21, 22, 23, 25, 26, 27, 40, 41, 48, 49, 51, 52, 53, 54, 55, 56, 58, 59, 62, 63, 64, 66, 67

BP Penalidade Por Brevidade - *Brevity Penality* 11, 12, 15, 17, 26, 59

BRAPT *Bilingual Rating of Psycholinguistic Perspectives in Translations* . . 5, 6, 27, 37, 38, 39, 40, 41, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 64, 65, 66, 67, 68

CA Computação Afetiva 1, 4, 7, 8

DARPA *Defense Advanced Research Projects Agency* 23

LIWC *Linguistic Inquiry Word Count* . . . 1, 5, 7, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 31, 37, 38, 40, 41, 43, 45, 46, 56, 65, 67, 68

MT Mineração De Textos 1, 4, 5, 7, 8, 21

NIST *National Institute of Standards and Technology* 23, 25, 26, 27

PB Português Do Brasil 2, 5, 37, 40, 49

PER *Position-independent Error Rate* 2, 23, 25, 26, 27

PLN Processamento De Linguagem Natural 1, 4, 8, 17

TAT Tradução Automática De Textos . . . 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 15, 17, 21, 23, 26, 27, 31, 36, 37, 38, 40, 41, 42, 43, 44, 45, 47, 48, 49, 51, 53, 62, 64, 65, 66, 67, 68

TD Tradução Direta 2, 15, 37

TI Tradução Inversa 2, 3, 37

WER *Word Error Rate* 2, 23, 24, 25, 26, 27

Capítulo I Introdução

Todos os dias surgem novas tecnologias visando substituir ou otimizar o trabalho humano e, cada vez mais, fazem parte do cotidiano das pessoas. Da mesma maneira, essas tecnologias precisam ser constantemente aprimoradas para que seus resultados sejam cada vez mais próximos daqueles alcançados pelo trabalho humano. Em áreas como a Computação Afetiva (CA) e a Mineração de textos (MT), alguns estudos analisam textos com o objetivo de realizar inferências acerca da idade [Rodrigues et al., 2017b] e do gênero do autor [da Ponte Junior et al., 2016]. Diversos estudos objetivam extrair aspectos linguísticos e psicológicos (*e.g.,* raiva, ansiedade, emoções positivas) a partir de textos [Guedes et al., 2016, Guimarães, 2016, Li et al., 2014, Rodrigues et al., 2016, Tavares et al., 2017]. Nesse sentido, diversas ferramentas têm sido utilizadas na tarefa de minerar e, principalmente, extrair características de textos, dentre as quais podemos citar, com grande destaque, o *Linguistic Inquiry Word Count* (LIWC) [Pennebaker et al., 2001a]. Essa ferramenta tem importância fundamental para o presente trabalho e por isso será abordada em maiores detalhes no momento oportuno.

Outra área de estudos, conhecida como Processamento de Linguagem Natural (PLN), é extremamente útil para essa tarefa de extrair padrões a partir de textos. Trata-se de um conjunto de técnicas computacionais utilizadas para analisar e representar textos em linguagem natural e que tem o propósito de auxiliar no processamento de idiomas para o desenvolvimento de diversas tarefas ou aplicações [Liddy, 2001]. Dentre essas tarefas é possível destacar a recuperação da informação, a geração automática de texto e a Tradução Automática de Textos (TAT) [Gelbukh et al., 2004].

As TATs são amplamente utilizadas e tornaram-se uma importante ferramenta no auxílio à comunicação no mundo globalizado, utilizando recursos do PLN para traduzir palavras e expressões de uma linguagem natural para outra, objetivando manter a equivalência semântica entre o conteúdo do texto original e do texto traduzido [Rodrigues et al., 2017a]. Dentre os diversos problemas que podem surgir do PLN, os principais decorrem de não-determinismos e de ambiguidades [da Rocha, 2007]. Nesses aspectos, por utilizar recursos do PLN, as TATs também estão sujeitas a ter sua qualidade comprometida.

Dessa maneira, a TAT ainda não substitui o trabalho de um tradutor profissional por completo, sendo necessária a revisão humana do texto traduzido, visto que esse processo ainda apresenta

dificuldades a serem sanadas, principalmente, no que diz respeito à consideração de semântica a partir dos textos traduzidos [Sales, 2011]. Essa equivalência semântica nem sempre é alcançada, podendo ocasionar traduções que diferem do conteúdo original (*e.g.,* apresentando perdas de significado e de aspectos linguísticos e psicológicos), quando comparadas com a tradução realizada por um especialista humano (*i.e.,* uma tradução referência) [Baskaya et al., 2017]. Essas dificuldades refletem uma necessidade de melhoria das ferramentas de TAT e também das métricas utilizadas atualmente para avaliar a qualidade desse tipo de tradução.

Existem, atualmente, diversas métricas para avaliar a qualidade das TATs (*e.g., Word Error Rate* (WER) e *Position-independent Error Rate* (PER)). Boa parte dessas métricas consiste em avaliar a compatibilidade entre uma tradução referência (*i.e.,* uma tradução confiável realizada por um humano especialista) e uma tradução candidata (*i.e.,* uma tradução cuja qualidade pretende-se avaliar). Dentre essas métricas, pode-se destacar a métrica *Bilingual Evaluation Understudy* (BLEU) [Papineni et al., 2002], que ainda permanece como o estado-da-arte [Chang et al., 2008, Zeng et al., 2014], sendo a mais utilizada dentre as métricas que avaliam a qualidade das TATs [Cer et al., 2010]. Nos exemplos a seguir é possível observar problemas decorrentes da TAT, bem como as dificuldades encontradas por métricas como a BLEU, concebidas para avaliar a qualidade dessas traduções.

I.1 Exemplos de traduções produzidas por ferramentas de TAT

Para evidenciar as dificuldades encontradas pelas ferramentas de TAT mencionadas no início deste capítulo, foram escolhidas três ferramentas bastante utilizadas em ambiente *web*: Google Tradutor[1], Bing Tradutor[2] e Babylon Tradutor[3]. Torna-se importante esclarecer, de antemão, que o objetivo do presente trabalho não é o de resolver os problemas das TATs. O objetivo principal é propor uma métrica mais eficaz, no sentido de considerar aspectos linguísticos e psicológicos, do que as métricas existentes para avaliar a qualidade desse tipo de tradução. No entanto, como missão precípua, faz-se necessário evidenciar os problemas das TATs para, então, estabelecer uma linha de entendimento que permita chegar aos problemas relacionados às métricas criadas para avaliá-las.

Voltando aos problemas das TATs, o estudo de Ferreira [2012] define como Tradução Direta (TD) a tradução de textos correlatos da língua estrangeira (*e.g.,* língua inglesa) para a língua materna (*e.g.,* língua portuguesa) e Tradução Inversa (TI) a tradução da língua materna (*e.g.,* língua portuguesa) para a língua estrangeira (*e.g.,* língua inglesa). Em relação às TDs, foram analisadas as traduções para o Português do Brasil (PB) da seguinte sentença em língua inglesa: “He’d been

[1] https://www.google.com.br/translator
[2] https://www.bing.com/translator/
[3] http://tradutor.babylon-software.com/

feeling blue all week.". As três ferramentas produziram, respectivamente, as seguintes traduções: "Ele estava sentindo azul toda a semana.", "Ele estava se sentindo azul durante toda a semana." e "Ele tinha sido sensação azul toda a semana.". É possível observar que as três ferramentas traduziram a sentença de maneira incorreta. Percebe-se, ainda, que as três sentenças apresentaram um resultado equivocado em relação à tradução da simples expressão "He'd been felling", principalmente no que diz respeito à flexão verbal. Em relação à tradução da palavra ambígua "blue", que, dependendo do contexto, pode significar uma cor (*i.e.,* azul) ou um estado de espírito (*e.g.,* triste, melancólico), é possível observar, novamente, traduções equivocadas produzidas pelas três ferramentas.

Da mesma maneira, ao realizar automaticamente uma TI para a língua inglesa, as ferramentas de TAT também podem gerar resultados incorretos. A sentença "Ela recebeu várias balas no dia das bruxas.", por exemplo, foi traduzida pelas três ferramentas mencionadas no parágrafo anterior. As referidas ferramentas produziram, respectivamente, as seguintes traduções: "She got several bullets on Halloween.", "She received several bullets on Halloween." e "She has received several bullets on Halloween.". Nesse caso também é possível observar problemas de tradução com relação à flexão verbal da expressão "Ela recebeu" e da palavra ambígua "bala".

I.2 Definição do problema

O estudo de Sales [2011] destaca que, apesar de tratar-se de uma área com mais de 60 anos de estudos, a confiabilidade das traduções produzidas por ferramentas de TAT ainda depende de revisão humana, principalmente para superar desafios que consistem na tradução de expressões idiomáticas, palavras ambíguas, verbos flexionados, dentre outros. Nesse panorama, o autor alerta para o fato de que tais ferramentas ainda não são capazes de substituir o trabalho humano, dado que, além dos aspectos já mencionados, não identificam precisamente o contexto em que palavras e expressões se encontram inseridas. Esse problema, ainda segundo o autor, existe em decorrência das traduções dependerem de conhecimentos gramaticais e léxicos, como também aspectos semânticos e pragmáticos referentes ao significado e ao uso da língua em contexto.

Conforme já mencionado, o objetivo das métricas que avaliam TATs é calcular a compatibilidade entre uma tradução referência e uma tradução candidata. No entanto, também são evidenciadas dificuldades na avaliação da qualidade desse tipo de tradução. Para ilustrar esse cenário, a métrica BLEU foi aplicada considerando a sentença referência "A minha casa é muito `bela`" e a sentença candidata "A minha casa é muito `bonita`". Embora estejamos comparando duas traduções muito parecidas, em que houve a simples substituição de uma palavra ("bela") por um sinônimo ("bonita"), de acordo com a BLEU, o percentual de compatibilidade entre as duas traduções foi calculado em apenas 53.73%. Isso ocorre por que essa métrica é incapaz de ana-

lisar as sentenças semanticamente e verificar que ambas as palavras são sinônimas e que, no contexto dessas sentenças, a utilização de uma ou outra não causa grande prejuízo à tradução.

A ideia central da BLEU é avaliar uma tradução candidata em comparação com uma tradução referência. 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 dessa natureza, por serem baseadas em pareamento exato e ordenado de palavras, não são capazes de considerar equivalentes sentenças diferentes, mas semanticamente semelhantes [de Melo et al., 2015]. Justamente por essa razão, o percentual de compatibilidade no exemplo da substituição de palavras sinônimas ficou tão baixo.

A maioria das métricas existentes e mais utilizadas consiste basicamente em verificar a correspondência de termos (ou palavras) contidos em textos como meras sequências de caracteres, limitando-se a verificar se tais sequências são iguais ou não, desconsiderando, portanto, a semântica desses textos ou termos. Nesse aspecto, estudos anteriores apontam para a importância e a necessidade de considerar a semântica das palavras na avaliação de TATs com o objetivo de melhorar a eficácia desse processo [Rodrigues et al., 2017a, Rodrigues and Guedes, 2017]. Torna-se, portanto, cada vez mais necessário o surgimento de novas métricas que sejam capazes de considerar semântica (*e.g.,* aspectos linguísticos (verbos, pronomes, advérbios) e psicológicos (afeto, ansiedade, emoções positivas)) no processo de avaliação de TATs, visando melhorar sua eficácia.

Dentre as diversas métricas existentes para avaliar a qualidade das TATs, conforme destacado no início deste capítulo, a métrica BLEU é a preferida e a mais utilizada [Cer et al., 2010, Santiago, 2013], além de permanecer como atual estado da arte, de acordo com os estudos de Callison-Burch et al. [2006] e Jones et al. [2009]. Nesse cenário, a BLEU foi escolhida para ser abordada com mais destaque no presente trabalho.

I.3 Objetivos

Esta dissertação se encontra contextualizada em áreas como Computação Afetiva (CA), Mineração de Textos (MT) e Processamento de Linguagem Natural (PLN), mais especificamente no âmbito da Tradução Automática de Textos (TAT) e possui como principais objetivos:

A Propor um novo algoritmo, denominado `calcDPC`, capaz de identificar em cada aspecto linguístico ou psicológico, o percentual de divergências na comparação entre uma tradução referência e uma tradução candidata. Dessa maneira, pode-se perceber que houve perda em relação a um determinado aspecto linguístico (*e.g.,* verbos, preposições) ou em relação a um determinado aspecto psicológico (*e.g.,* emoções positivas, afeto).

B Propor uma nova métrica, denominada *Bilingual Rating of Psycholinguistic Perspectives in Translations* (BRAPT), capaz de avaliar aspectos psicológicos e linguísticos, ao envolver informação semântica em traduções produzidas por ferramentas de TAT. Para incorporar a parte semântica, a métrica proposta utiliza um léxico afetivo para representar as traduções referência e candidata em forma de vetores, além de uma técnica conhecida na área de MT para verificar a compatibilidade entre esses vetores.

C Utilizar textos traduzidos para o PB por especialistas humanos e por ferramentas de TAT com o intuito de comparar a métrica proposta (*i.e.,* BRAPT) ao estado da arte (a métrica BLEU) na avaliação dessas TATs; verificar as vantagens e, posteriormente, algumas desvantagens da nova métrica em relação ao estado da arte.

D Avaliar três ferramentas de TAT conhecidas, disponíveis em ambiente *web* mostrando em percentuais o desempenho de cada uma delas utilizando as métricas BLEU e BRAPT.

E Realizar uma análise detalhada de sentenças em que o estado da arte (*i.e.,* a métrica BLEU) se mostrou ineficaz. O objetivo é utilizar as duas métricas (*i.e.,* BLEU e BRAPT) para analisar aspectos como: a substituição de palavras por sinônimos; inversão de ordem de palavras que não causa impacto significativo na compreensão da mensagem a ser transmitida; diminuição no tamanho da sentença candidata ao substituir dois termos por um sinônimo (*e.g.,* “muito bonita” por “linda”), dentre outros.

A métrica proposta também deve se mostrar capaz de avaliar TATs nos aspectos em que o estado da arte (*i.e.,* a métrica BLEU) se mostrou ineficaz, ou seja, na observação da semântica (*e.g.,* aspectos linguísticos e psicológicos) contida nesse tipo de tradução. Para esse propósito, a métrica proposta utiliza uma ferramenta de análise textual e processamento de linguagem natural denominada LIWC. Essa ferramenta possui versões correlatas para a língua inglesa e para o português do Brasil (PB) e tem a capacidade de rotular palavras em categorias que refletem aspectos linguísticos e psicológicos, gerando, para cada sentença de um texto, um vetor com as posições correspondentes a essas categorias.

I.4 Organização da dissertação

Esta dissertação está organizada em mais seis capítulos. O capítulo II fornece todos os conhecimentos necessários para a compreensão do assunto abordado por meio de conceitos e explicações objetivas sobre o problema a ser resolvido, as motivações, contribuições e áreas envolvidas neste estudo. O capítulo III descreve alguns trabalhos relacionados a métricas para avaliação de traduções automáticas de textos. O capítulo IV trata da identificação do percen-

tual de representatividade de cada aspecto linguístico ou psicológico presente em uma determinada sentença, além do algoritmo `calcDPC`, capaz de identificar as divergências em cada aspecto linguístico ou psicológico verificado entre duas sentenças. O capítulo V discorre sobre a contribuição principal desta dissertação: a métrica BRAPT, a metodologia adotada e o algoritmo proposto para realizar o cômputo de sua compatibilidade. O capítulo VI apresenta as avaliações experimentais realizadas por meio da análise de dez textos com 128 sentenças traduzidas por dois humanos e por três ferramentas de TAT. Por fim, o capítulo VII traz as considerações sobre os resultados obtidos, bem como dificuldades e limitações encontradas, além de apresentar as conclusões acerca das avaliações experimentais, cenários futuros e possíveis contribuições.

Capítulo II Fundamentação teórica

Neste capítulo são apresentadas as áreas envolvidas neste estudo, além dos conceitos essenciais para o entendimento do presente trabalho. Esses conceitos são fundamentais para a compreensão do conteúdo dos capítulos seguintes. Os conceitos estão divididos em seis seções: a Seção II.1 descreve a área conhecida como Computação Afetiva (CA) e suas aplicações; a Seção II.2 discorre sobre a Mineração de Textos (MT) e a necessidade de utilização dessa técnica para obter os resultados esperados no presente trabalho; a Seção II.3 aborda conceitos referentes à Tradução Automática de Textos (TAT), tema principal deste trabalho, e suas características; a subseção II.3.1 apresenta os desafios a serem superados pela TAT, a subseção II.3.2 discorre sobre como as TATs costumam ser avaliadas e a subseção II.3.3 apresenta as limitações enfrentadas na tarefa de avaliar TATs e, consequentemente, as motivações para a proposição de novas métricas que sejam capazes da avaliar TATs a partir de aspectos semânticos (*e.g.,* aspectos linguísticos e psicológicos); a seção II.4 descreve a métrica BLEU que é considerada o estado da arte e a mais utilizada dentre as métricas existentes e, portanto, utilizada como referência para o presente trabalho. A formulação dessa métrica é apresentada na subseção II.4.1; a subseção II.4.2 traz a aplicação da fórmula utilizada para calcular e métrica BLEU e a subseção II.4.3 traz, a partir do exemplo de sua utilização, algumas limitações dessa métrica; a Seção II.5 discorre sobre a ferramenta LIWC e sua essencial contribuição para essa tarefa de avaliar TATs, observando os aspectos linguísticos e psicológicos necessários à métrica proposta no presente trabalho. Por fim, a seção II.6 aborda a medida de similaridade entre textos conhecida como Similaridade do Cosseno. A nova métrica proposta no presente trabalho utiliza essa medida para verificar a similaridade entre vetores que representam as sentenças a serem analisadas.

II.1 Computação afetiva

A computação afetiva (CA) é uma área de estudos introduzida por Rosalind Wright Picard em 1995. Trata-se de uma área multidisciplinar que utiliza diversos campos do conhecimento (*e.g.,* Informática, Educação, Psicologia, Sociologia, Inteligência Artificial, dentre outros) para criar ferramentas que sejam capazes de identificar aspectos emocionais do ser humano e tentar fazer com que máquinas simulem essas habilidades emocionais, resultando em sistemas afetivos. A

ideia é que esses sistemas afetivos sejam capazes de identificar e simular habilidades emocionais, tornando, portanto, a interação homem-máquina mais bem sucedida [Picard, 1995]. A partir desse cenário podem surgir inúmeras aplicações e estudos como a inferência de fatores da personalidade [Golbeck et al., 2011] e o desenvolvimento de robôs emocionais para auxiliar no tratamento de crianças doentes, por exemplo [Aly and Tapus, 2013].

Por tratar-se de uma área com pouco mais de 20 anos de estudos e, portanto, considerada recente, a CA ainda oferece muitas possibilidades a serem exploradas. Com os fundamentos dessa área é possível construir *softwares* e *hardwares* que facilitem o processo de identificação de aspectos afetivos ligados ao comportamento e às emoções humanas. Alguns estudos afirmam que essa identificação pode se dar por meio do reconhecimento do tom de voz, da linguagem corporal e da expressão facial [França et al., 2012, Picard, 1997].

Conforme mencionado, alguns estudos visam extrair características humanas e padrões a partir de textos [D'Mello and Graesser, 2012, Dzindolet and Pierce, 2005, Tatai and Laufer, 2004]. Estudos nessa direção possibilitam, inclusive, a identificação de patologias como a depressão a partir de textos escritos em mídias sociais [De Choudhury et al., 2013]. Esses estudos foram desenvolvidos com o auxílio de métodos provenientes da área de Mineração de Textos (MT).

II.2 Mineração de textos

A Mineração de Textos (MT), também conhecida como descoberta de conhecimento a partir de textos, surgiu em 1995 e foi mencionada pela primeira vez no trabalho de Feldman and Dagan [1995]. A MT consiste em analisar textos a partir de recursos computacionais com o objetivo de descobrir informações, características e identificar padrões a partir dos mesmos. Trata-se, portanto, de uma área de estudos que utiliza o Processamento de Linguagem Natural (PLN) e cujo objetivo é a recuperação e extração de informações [Feldman and Dagan, 1995].

Na MT há diversos estudos focados em extrair padrões e informações úteis a partir de textos [Hotho et al., 2005, Pereira et al., 2013, Schardong et al., 2013]. A MT pode ser tratada como um processo de especialização de uma área mais abrangente conhecida como Mineração de Dados. Esse processo consiste em cinco etapas que podem ser instanciadas de acordo com os objetivos de um projeto. São elas: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Uso do Conhecimento [Rezende et al., 2003]. O ciclo formado por essas etapas é ilustrado na figura II.1.

Atualmente, a MT tem se tornado bastante útil em estudos provenientes de muitas outras áreas do conhecimento (*e.g.,* Computação afetiva, análise de sentimentos, traduções de textos) em que há a necessidade de resolver problemas de representação de texto, extração de informações, classificação, agrupamento ou a busca e modelagem de padrões ocultos. Mui-

tos desses aspectos (*e.g.,* busca de padrões, extração de características) se fazem necessários para os estudos concentrados na área de Tradução Automática de Textos (TAT). Na fase de pré-processamentos, por exemplo, é realizada uma limpeza no texto a fim de eliminar pontuações, caracteres indesejáveis, padronizar para letras minúsculas, entre outros. Já na fase de extração de padrões, busca-se extrair aspectos linguísticos e psicológicos presentes nas traduções. Por fim, vem a utilização do conhecimento, onde é possível realizar análises de traduções com base em suas características linguísticas e psicológicas identificadas na fase anterior.

Figura II.1: Etapas do processo de mineração de textos. Fonte: [Rezende et al., 2003].

II.3 Tradução automática de textos

A Tradução Automática de Textos (TAT) utiliza a computação para converter uma mensagem de uma linguagem natural para outra, tentando manter a equivalência com o conteúdo original. Os primeiros projetos nessa área foram desenvolvidos durante a guerra fria, com motivações militares. Desde então, a TAT tornou-se um facilitador da comunicação na era globalizada [de Melo et al., 2015, Sales, 2011]. Atualmente, as ferramentas de TAT, especialmente aquelas que funcionam em ambiente *web*, estão cada vez mais presentes em nosso cotidiano. Essas ferramentas fazem uso de algumas técnicas como, por exemplo, as memórias de tradução; essa técnica, utilizada por algumas ferramentas, consiste em manter um grande banco de dados de traduções anteriores a serem utilizadas de acordo com o reaparecimento de elementos idênticos ou muito parecidos, configurando economia significativa de tempo no processo de tradução [Baskaya et al., 2017, Weininger, 2004].

II.3.1 Desafios a serem superados pelas TATs

O estudo de Sales [2011] destaca que, apesar de tratar-se de uma área com muitos anos de estudos, a confiabilidade desse tipo de tradução ainda depende de revisão humana, principalmente para superar desafios que consistem em problemas na tradução de expressões idiomáticas (*e.g., "To cost an arm and a leg", "Let the cat out of the bag", "To feel under the weather"*) e até mesmo expressões simples, na flexão verbal, na tradução de palavras ambíguas (*e.g., blue, interest, rare*) e na identificação do contexto em que essas palavras e expressões encontram-se inseridos. Alguns desses problemas já foram evidenciados nas seções I.1 e I.2.

No entanto, a questão principal a ser abordada no presente estudo não está concentrada nos problemas das TATs. A questão principal está concentrada justamente nas métricas existentes para avaliar esse tipo de tradução, uma vez que estas têm se mostrado ineficazes no sentido de detectar os problemas citados que, em geral, dizem respeito à questões de ordem semântica.

II.3.2 A avaliação de TATs

Existem diversas métricas para avaliar a qualidade das TATs. Em seu estudo, de Melo et al. [2015] destacam que essa avaliação, em geral, consiste em comparar uma tradução referência (*i.e.,* uma tradução confiável realizada por um humano especialista) com uma tradução cuja qualidade se pretende avaliar, também conhecida como tradução candidata. O objetivo dessas métricas é avaliar a compatibilidade entre essas traduções por meio de uma abordagem quantitativa, baseada no pareamento exato e ordenado das palavras. Ainda de acordo com o referido estudo, essas métricas geralmente limitam-se a comparar a correspondência dos termos (ou palavras) contidos nas sentenças avaliadas, verificando se as mesmas são iguais ou não.

A avaliação é feita de forma muito parecida na maioria das métricas. A verificação acerca da correspondência dos termos consiste em avaliar a precisão das sequências de *n* letras ou palavras, denominadas *n-gramas*. Um *3-grama* (*i.e.,* trigrama), por exemplo, reflete todas as combinações sequenciais possíveis de três termos para uma determinada sentença. A precisão de *n-gramas* indica a quantidade de *n-gramas* compatíveis entre a sentença candidata e a sentença referência, utilizando como base a quantidade total de palavras da sentença referência. Com o intuito de proporcionar melhor entendimento, podemos observar a sentença "A casa está bem conservada" e *n-gramas* variando entre 1 e 4:

- ***1-gramas*** (*i.e.,* unigramas): "A", "casa", "está", "bem", "conservada".
- ***2-gramas*** (*i.e.,* bigramas): "A casa", "casa está", "está bem", "bem conservada".
- ***3-gramas*** (*i.e.,* trigramas): "A casa está", "casa está bem", "está bem conservada".

- ***4-gramas*** (*i.e.,* tetragramas): “A casa está bem”, “casa está bem conservada”.

A expectativa dessas métricas é de que a sentença candidata tenha o mesmo ordenamento e número de palavras da sentença referência. Por isso, algumas dessas métricas também penalizam sentenças candidatas menores do que a sentença referência, em número de palavras. Esse tipo de penalidade é conhecida como Penalidade por Brevidade - *Brevity Penality* (BP) e é adotada por métricas muito utilizadas como a já citada BLEU e a NIST, que é abordada em maiores detalhes no capítulo III.

II.3.3 Dificuldades para avaliar TATs

O estudo presente em [de Melo et al., 2015] descreve a avaliação das TATs por meio de uma análise comparativa entre uma sentença referência e uma sentença candidata considerando a equivalência entre elas, utilizando, em boa parte das métricas, a abordagem por meio da comparação da equivalência dos já citados *n-gramas*. Os autores afirmam que esse tipo de avaliação não considera o teor das palavras e sentenças, resumindo-se a uma análise do ponto de vista métrico, como meras sequências de caracteres cujo significado é desconsiderado. De acordo com o referido estudo, tais métricas não são capazes de avaliar a qualidade das TATs de forma satisfatória. Essa afirmação é corroborada pelo estudo de Sales [2011], que acrescenta que esse tipo de avaliação depende de conhecimentos empíricos acerca do significado das palavras, observância de aspectos linguísticos, psicológicos, semânticos e de contexto; conhecimentos esses, peculiares aos seres humanos. Por fim, com base nessas conclusões, de Melo et al. [2015] afirmam que as métricas existentes ainda se mostram incapazes de substituir o trabalho humano.

II.4 A métrica BLEU

A métrica BLEU foi desenvolvida pela IBM e, como já foi mencionado, em linhas gerais essa métrica compara uma sentença candidata (traduzida por uma ferramenta de TAT) com uma sentença referência (traduzida por um especialista humano) com o objetivo de verificar a compatibilidade entre essas duas traduções [Papineni et al., 2002]. Essa métrica apresenta valores entre 0 e 1. No entanto, para o presente trabalho optou-se por apresentar os resultados em percentuais. Portanto, deste ponto em diante, essa será a abordagem adotada até o fim desta dissertação.

Por ser a mais popular, a mais utilizada e por se tratar do atual estado da arte em relação a métricas utilizadas para avaliar a qualidade de TATs, a métrica BLEU e o entendimento de seu funcionamento são de fundamental importância para o presente trabalho. Por essa razão,

optou-se por dedicar uma subseção exclusivamente para explicar, de forma bem detalhada, toda a formulação que envolve essa métrica.

II.4.1 A formulação da métrica BLEU

A métrica BLEU indica a quantidade de *n-gramas* compatíveis entre as referidas sentenças, considerando a quantidade total de palavras de ambas as sentenças. Essa quantidade de *n-gramas* compatíveis recebe o nome de precisão modificada de *n-gramas*. Ao comparar uma sentença referência com uma sentença candidata com o objetivo de verificar a similitude entre as traduções, a métrica BLEU considera a precisão modificada de *n-gramas*, com o parâmetro *n* assumindo, em sua versão clássica, valores entre 1 (*i.e.,* unigramas) e 4 (*i.e.,* tetragramas) [Papineni et al., 2002]. A precisão modificada de *n-gramas* é representada por p_i, onde i refere-se ao nível de *n-grama* utilizado. Dessa forma, nesse contexto, essa precisão pode variar entre p_1 e p_4.

O objetivo da BLEU é verificar o quão próximo o resultado produzido por um tradutor automático chega do resultado produzido por um especialista humano, calculando a precisão modificada dos *n-gramas* na sentença candidata (*i.e.,* a sentença que se quer avaliar) em comparação com a sentença referência (*i.e.,* a sentença tida como confiável). Esse objetivo consiste em 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. Essa métrica baseia-se na suposição de que uma boa tradução tem mais *n-gramas* compatíveis com a sentença referência do que uma tradução ruim [Finch et al., 2004].

Quando a sentença candidata é menor do que a sentença referência, atribui-se uma penalidade denominada Penalidade por Brevidade (BP), pois, de acordo com essa métrica, a sentença candidata deve ser semelhante à sentença referência em tamanho, escolha e ordem de palavras. A BP deve ser contabilizada uma única vez e não a cada sequência de *n-gramas*. Na Equação 1 exibida abaixo, podemos observar o cálculo da BP, em que r refere-se à quantidade de palavras da sentença referência e c representa a quantidade de palavras presentes na sentença candidata.

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

A métrica BLEU, exibida na Equação 2, pode ser obtida calculando-se a média geométrica das precisões modificadas p_i, referentes a *n-gramas* cujos níveis variam de 1 até n e pesos positivos w_i vinculados aos seus respectivos *n-gramas*. Tudo isso multiplicado pela BP. Nesse caso, se o resultado da BP for obtido por $e^{(1-\frac{r}{c})}$, sempre vai apresentar valores menores que 1 e por se tratar

de um fator multiplicador, consequentemente, faz com que o cômputo do valor da métrica BLEU seja calculado para baixo.

$$BLEU = BP \cdot e^{(\Sigma_{i=1}^{n} w_i \log p_i)} \tag{2}$$

As equações e explicações aqui apresentadas a respeito da métrica de BLEU constam no estudo de Papineni et al. [2002]. Como já foi mencionado, no *baseline* do referido estudo, foi utilizado o limite 4 para *n* e pesos uniformes w_i. Desta forma, os pesos dependem da variável n e podem ser denominados w, conforme pode ser observado na Equação 3,

$$BLEU = BP \cdot e^{(w \Sigma_{i=1}^{n} \log p_i)} \tag{3}$$

Como o somatório de vários logs corresponde ao log de vários produtos, é possível simplificar novamente, conforme ilustrado na Equação 4.

$$BLEU = BP \cdot e^{(w. \log(\Pi_{i=1}^{n} p_i))} \tag{4}$$

Como um fator multiplicando um log corresponde a um log elevado a esse fator, tem-se, na Equação 5, a seguinte representação:

$$BLEU = BP \cdot e^{(\log(\Pi_{i=1}^{n} p_i)^w)} \tag{5}$$

Log e exponenciais são funções inversas. A exponencial do log de um valor é exatamente igual a este valor. Logo, a Equação 6 é representada de forma ainda mais simplificada.

$$BLEU = BP \cdot \left(\prod_{i=1}^{n} p_i \right)^w \tag{6}$$

A aplicação dessas operações algébricas decorrentes das propriedades das funções exponenciais e logarítmicas são baseadas no trabalho de dos Santos Machado [1995]. A aplicação dessas operações objetivou facilitar o entendimento da equação que da origem à métrica BLEU. Formalmente a pontuação da referida métrica pode ser obtida por meio da Equação 7.

$$BLEU = BP \cdot (\prod_{i=1}^{n} p_i)^{\frac{1}{n}} \tag{7}$$

Lembrando que, nesse caso, n = 4, a métrica BLEU pode ser exemplificada, de forma mais simplificada ainda na Equação 8.

$$BLEU = BP \cdot (p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{0.25} \tag{8}$$

II.4.2 Exemplo de aplicação da métrica BLEU

O estudo de Koehn [2009] ilustra um exemplo de sentenças com as identificações dos *n-gramas* por meio da figura II.2, em que é possível observar que a sentença candidata representada por SYSTEM A possui um *2-grama* e um *1-grama*. Da mesma forma, percebe-se que a sentença candidata representada por SYSTEM B possui um *2-grama* e um *4-grama*. A sentença referência possui sete palavras e cada uma das duas sentenças candidatas possui seis palavras. Nas duas sentenças candidatas percebemos a marcação de três dos quatro tipos de *n-gramas* (*i.e.,* unigramas, bigramas e tetragramas), lembrando que os bigramas consistem em todas as combinações sequenciais possíveis de duas palavras em uma sentença, os trigramas consistem em todas as combinações sequenciais possíveis de três palavras e que os tetragramas consistem em todas as combinações sequenciais possíveis de quatro palavras. Já os unigramas são todas as palavras presentes na referida sentença.

Figura II.2: Aplicação da métrica BLEU com 2 candidatas. Fonte [Koehn, 2009].

Ao observar SYSTEM B, é de fácil percepção que um trecho com um *4-grama* de correlação de padrão, como é o caso do trecho “Israeli officials are responsible” contém dois trechos de *3-grama* (*i.e.,* “israelli officials are” e “officials are responsible”), três trechos de *2-grama* (*i.e.,* “israeli officials”, “officials are” e “are responsible”) e, ainda, quatro trechos de *1-grama* (*i.e.,* “israeli”, “officials”, “are” e “responsible”). A métrica BLEU foi aplicada com base no exemplo ilustrado na figura II.2, como pode ser visto na tabela II.1 e nos cálculos subsequentes.

Tabela II.1: Aplicação da métrica BLEU.

Métrica	**System A**	**System B**
precisão (*1-gramas*)	$3/6$	$6/6$
precisão (*2-gramas*)	$1/5$	$4/5$
precisão (*3-gramas*)	$0/4$	$2/4$
precisão (*4-gramas*)	$0/3$	$1/3$
penalidade por brevidade	$e^{(1-7/6)}$	$e^{(1-7/6)}$
BLEU	0%	51%

Nesse caso, como a quantidade de palavras (*i.e.,* termos) contidas nas sentenças candidatas

SYSTEM A e SYSTEM B são inferiores à quantidade de palavras contidas na sentença referência, a BP foi aplicada para ambas as sentenças. Na referida figura também é possível verificar a precisão modificada de *n-gramas* em cada candidata, bem como a penalidade por brevidade.

Também se faz importante observar que a candidata SYSTEM A, por apresentar incompatibilidade e, consequentemente, precisão modificada de *n-gramas* zerada em relação aos *3-gramas* (trigramas) e *4-gramas* (tetragramas), ficou com a compatibilidade em 0.0%, ou seja, incompatibilidade total. Sempre que ao menos um dos *n-gramas* apresentar precisão igual a 0, automaticamente a compatibilidade BLEU também vai ser igual a 0. Considerando a sentença referência representada por $SRef$ e a sentença candidata SYSTEM B representada por $SysB$, com $n = 4$, a figura II.3 ilustra, passo a passo, o cálculo da compatibilidade BLEU.

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

$$BLEU(SRef, SysB) = e^{\left(1-\frac{7}{6}\right)} \cdot (1 \cdot 0,8 \cdot 0,5 \cdot 0,333333)^{0,25}$$

$$BLEU(SRef, SysB) = e^{\left(1-\frac{7}{6}\right)} \cdot (0,133333)^{0,25}$$

$$BLEU(SRef, SysB) = 0,846482 \cdot 0,604275$$

$$BLEU(SRef, SysB) = 0,51151$$

$$BLEU(SRef, SysB) = 0,51151 \cdot 100 = 51,151\%$$

Figura II.3: Exemplo de aplicação da BLEU à candidata System B.

Cabe salientar que tanto as ilustrações quanto os cálculos apresentados acima, constam no estudo de [Koehn, 2009] e que para simplificar o entendimento, houve apenas um nível de detalhamento maior dos cálculos, que foram representados em diversas etapas, para que se pudesse notar cada mudança.

Diante da importância da métrica BLEU para o presente trabalho, tornou-se necessário facilitar ao máximo seu entendimento por meio de exemplos, ilustrações e aplicação da equação utilizada para calcular sua compatibilidade.

II.4.3 Limitações da métrica BLEU

Apesar da TAT estar sendo estudada há mais de meio século, de Melo et al. [2015] alertam que essa ainda é uma área pouco desenvolvida e explorada e relatam que métricas como a BLEU ainda são insuficientes para avaliar a qualidade das TATs. A partir de uma tradução referência de uma sentença em inglês para o português do Brasil (*i.e.,* uma tradução direta (TD)) realizada por

um especialista humano e uma tradução candidata da mesma sentença, realizada por uma das ferramentas de TAT citadas (*i.e.,* Google Tradutor, Bing Tradutor e Babylon Tradutor), percebe-se que a métrica BLEU, por exemplo, mostra um rigor muito grande em relação à substituição de apenas um único termo (ou palavra) entre essas sentenças.

A figura II.4 indica, em cada *n-grama* destacado em amarelo, as incompatibilidades verificadas em relação à substituição desse único termo na comparação entre a sentença referência "Ele tem algumas reuniões `com` clientes." e a sentença candidata B "Ele tem algumas reuniões `de` clientes". Essas incompatibilidades geram punições em relação à precisão de cada *n-grama* (*i.e.,* sequências de *n* letras ou palavras, conforme já citado na subseção II.4.1). A precisão modificada de *n-gramas* está representada na referida figura pela coluna intitulada "p-n". A precisão modificada de cada *n-grama* varia entre zero e um. Quanto maior for o valor dessa precisão maior será a compatibilidade em relação aos *n-gramas* analisados.

n-gramas	sentenças	Compatibilidade BLEU (Referência vs Candidata B)							p-n
1-grama	Referência	Ele	tem	algumas	reuniões	com	clientes		5/6
	Cand. B	Ele	tem	algumas	reuniões	de	clientes		
2-grama	Referência	Ele tem	tem algumas	algumas reuniões	reuniões com	com clientes			3/5
	Cand. B	Ele tem	tem algumas	algumas reuniões	reuniões de	de clientes			
3-grama	Referência	Ele tem algumas	tem algumas reuniões	algumas reuniões com	reuniões com clientes				2/4
	Cand. B	Ele tem algumas	tem algumas reuniões	algumas reuniões de	reuniões de clientes				
4-grama	Referência	Ele tem algumas reuniões	tem algumas reuniões com	algumas reuniões com clientes					1/3
	Cand. B	Ele tem algumas reuniões	tem algumas reuniões de	algumas reuniões de clientes					

Figura II.4: Comparação entre uma sentença candidata e uma sentença referência pela métrica BLEU.

Ainda sobre a figura II.4, cabe observar que nos casos em que se configura incompatibilidade, os valores das precisões modificadas de *n-gramas* tendem a diminuir gradativamente conforme aumenta-se o grau do *n-grama* analisado. Isso significa que a precisão verificada nos *4-gramas* (*i.e.,* tetragramas) tende a ser menor do que a precisão verificada nos *3-gramas* (*i.e.,* trigramas). A precisão verificada nos *3-gramas* tende a ser menor que a precisão verificada nos *2-gramas* (*i.e.,* bigramas) e assim por diante, até chegar aos *1-gramas* (*i.e.,* unigramas). No exemplo abordado, de acordo com a métrica BLEU, o percentual de compatibilidade da sentença candidata (*i.e.,* Candidata B) em relação à sentença referência é de 53.73%.

Apesar de não representar garantia de compatibilidade, palavras pertencentes a uma mesma categoria linguística muitas vezes ajudam a diminuir o impacto causado pela substituição. De todo modo, também não se pode deixar de mencionar que palavras de uma mesma categoria linguística muitas vezes podem significar até mesmo situações opostas. Tomemos como exem-

plo as preposições “para” e “da” e sua aplicação nas seguintes sentenças: “Ele veio `para` a minha casa.” e “Ele veio `da` minha casa.”. Nesse caso, por exemplo, as preposições “para” e “da” representam, respectivamente, destino e origem. Isso impacta significativamente no entendimento da mensagem que se deseja transmitir. Por isso, se faz necessário fazer essa ressalva.

A equação utilizada para chegar ao percentual de compatibilidade da métrica BLEU foi detalhada na subseção II.4.1 e mostra que, no caso de candidatas menores, considera-se uma penalidade, conhecida como BP, multiplicada pelo produtório das precisões modificadas de *n-gramas*, que por sua vez é elevado $1/n$, que nesse caso é 0.25. No exemplo ilustrado na figura II.4, como ambas as sentenças são de mesmo tamanho, não houve penalização da Candidata B ($CandB$) nesse sentido e chegou-se ao percentual de 53.73% de compatibilidade por meio do seguinte cálculo, em que 1, presente logo após o sinal de “=”, representa a ausência de uma BP.

$$BLEU(SRef, CandB) = 1 \cdot (0,83 \cdot 0,6 \cdot 0,5 \cdot 0,333333)^{0,25} = 0,5373$$

$$BLEU(SRef, CandB) = 53,73$$

Torna-se importante destacar que mesmo com a ausência de uma penalidade como a BP, a substituição de um único termo impacta consideravelmente no percentual de compatibilidade de acordo com a métrica BLEU, ainda que, como no exemplo abordado, o comprometimento da compreensão da sentença candidata B em relação à sentença referência seja pequeno. Nessa situação específica, os termos envolvidos pertencem à mesma categoria linguística (*i.e.,* a categoria das preposições).

Outro fato importante a ser considerado é que mesmo que uma palavra contida na sentença referência fosse substituída por um sinônimo na sentença candidata, o resultado teria sido o mesmo, visto que a BLEU não observa aspectos linguísticos e psicológicos dos termos analisados. Métricas como a BLEU são incapazes de verificar aspectos semânticos que podem ser facilmente observados em uma avaliação humana. Portanto, a possibilidade de extrair características que auxiliem na identificação de semântica (*e.g.,* características linguísticas e psicológicas) em textos representa um ganho extremamente relevante nesse processo. Léxicos afetivos como o LIWC possuem essa capacidade e podem auxiliar a adicionar semântica ao processo de avaliação de TATs.

II.5 LIWC

O LIWC é uma ferramenta de análise textual e Processamento de Linguagem Natural (PLN) amplamente utilizada na busca e identificação de padrões e na extração de aspectos linguísticos, psicológicos e sociais que podem ser utilizados em estudos que envolvam as mais diversas áreas

do conhecimento. Essa ferramenta é capaz de analisar os componentes emocionais, cognitivos e estruturais de textos. Isso é feito por meio de um léxico de termos (*i.e.*, palavras e pontuações) que pode ser encontrado em diversos idiomas [Pennebaker et al., 2007]. A primeira versão do LIWC foi proposta para a lingua inglesa em 2001 [Pennebaker et al., 2001b]. A versão de 2015 do LIWC em língua inglesa conta com cerca de 4.500 termos distribuídos em 80 categorias, sendo 68 categorias de palavras e 12 categorias de pontuações.

O trabalho de Tausczik and Pennebaker [2010] revisa vários métodos de análise de texto computadorizado e descreve como o LIWC foi criado e validado. Trata-se de um programa de análise de texto transparente que conta palavras em categorias significativas tanto do ponto de vista linguístico (*e.g.,* pronomes pessoais, verbos, advérbios, preposições, conjunções) quanto do ponto de vista psicológico (*e.g.,* afeto, emoções positivas, emoções negativas, ansiedade, aspectos cognitivos). O LIWC se mostrou capaz de detectar significado em uma ampla variedade de cenários experimentais, como por exemplo, mostrar foco de atenção, emoções, relações sociais, estilos de pensamento e diferenças individuais.

Esses estudos - que duraram muitos anos e envolveram profissionais de diversas áreas - evidenciaram a existência de duas categorias muito amplas de palavras que têm propriedades psicométricas (*i.e.,* que visam a medição dos fenômenos psíquicos) e psicológicas diferentes: as palavras de conteúdo e as palavras de estilo (também conhecidas como palavras de função). As palavras de conteúdo, em geral, são substantivos, verbos regulares, muitos adjetivos e advérbios; elas transmitem o conteúdo de uma comunicação. Usando como exemplo a frase “Foi uma noite escura e tempestuosa”, as palavras de conteúdo são: “escura”, “tempestuosa” e “noite”. Junto às palavras de conteúdo estão as palavras de estilo (ou de função). As palavras de estilo são constituídas por pronomes, preposições, artigos, conjunções, verbos auxiliares e algumas outras categorias. Na referida frase, essas palavras são: “foi”, “uma”, “e” [Tausczik and Pennebaker, 2010].

Miller [1991] ressalta que embora existam quase 100.000 palavras no vocabulário inglês, apenas cerca de 500 (*i.e.,* $0,05\%$) são palavras de estilo/função. Entretanto, essas palavras compõem cerca de 55% das palavras utilizadas na comunicação de um modo geral, seja ela escrita ou falada. Ainda segundo o autor, nosso cérebro tende a processar palavras de conteúdo e palavras de estilo/função de forma muito diferente. As palavras de conteúdo revelam o que se quer comunicar enquanto as palavras de estilo/função revelam a forma e a maneira particular como cada indivíduo se comunica, além de revelar mais sobre aspectos psicológicos e emocionais na análise de mensagens e textos.

O LIWC possui uma versão para o português do Brasil que conta com 127.149 palavras, cada uma relacionada a uma ou mais categorias. Essas categorias representam perspectivas

linguísticas, psicológicas, dentre outras. O dicionário é composto por um total de 64 categorias (*e.g., posemo* (*i.e.,* emoções positivas), *pronoun*, *time* (*i.e.,* aspectos relacionados ao tempo)). Cada palavra presente em um texto pode ser expressa por diferentes categorias que, por sua vez, refletem diferentes perspectivas [Filho, 2013].

Nas versões de 2007 do LIWC para o português do Brasil e para a língua inglesa, as 64 categorias podem ser descritas como sendo 27 categorias principais e 37 subcategorias. As categorias *anxiety (anx), anger (ang) e sadness (sad)* (*i.e.,* ansiedade, raiva e tristeza), por exemplo, são subcategorias de *negative emotion (negemo)* (*i.e.,* emoções negativas) e as categorias *personal pronouns (ppron)* (*i.e.,* pronomes pessoais) e *indefinite pronouns (ipron)* (*i.e.,* pronomes indefinidos) são subcategorias de *pronouns (pronoun)* (*i.e.,* pronomes).

Cada categoria do LIWC é representada por uma numeração não sequencial e uma nomenclatura em língua inglesa, condizente com sua finalidade. A figura II.5 ilustra essas categorias e os aspectos que cada uma reflete nas versões de 2007 em língua inglesa e PB, por exemplo.

	Algumas das 64 categorias do LIWC				
Aspectos linguísticos	3-ppron	11-verb	16-adverb	17-prep	18-conj
Aspectos psicológicos	125-affect	126-posemo	127-negemo	128-anx	131-cogmech

Figura II.5: Algumas das categorias do LIWC que refletem aspectos linguísticos e psicológicos.

A tabela II.2 traz alguns exemplos de palavras classificadas em categorias que representam aspectos linguísticos e categorias que representam aspectos psicológicos.

Tabela II.2: Palavras classificadas em categorias do LIWC

Categoria	Descrição	Aspecto	Palavras
16-adverb	Advérbios	Linguístico	“acolá”, “aonde”, “bastante”, “muito”
3-ppron	Pronomes pessoais	Linguístico	“dele”, “eu”, “me”, “nossa”
17-prep	Preposições	Linguístico	“até”, “com”, “de”, “para”
11-verb	Verbos	Linguístico	“abrir”, “cantam”, “desconfiou”, “vendem”
128-anx	Ansiedade	Psicológico	“desconfiou”, “inquietude”, “oprimirem”
131-cogmech	Cognição	Psicológico	“ideia”, “identico”, “senti”, “vi”
127-negemo	Emoções negativas	Psicológico	“aterrorizados”, “desconfiou”, “odiar”
126-posemo	Emoções positivas	Psicológico	“amor”, “brincamos”, “feliz”, “virtude”

II.5.1 Representação vetorial do LIWC

Como já foi mencionado, no LIWC uma palavra pode estar relacionada a diversas categorias e cada categoria abrange diversas palavras. A figura II.6 ilustra uma sentença com suas palavras contabilizadas em um vetor com as 64 categorias do LIWC em que é possível observar que a palavra “costuma”, por exemplo, foi contabilizada em 3 categorias e que a categoria representada pela posição 7 teve 3 palavras contabilizadas.

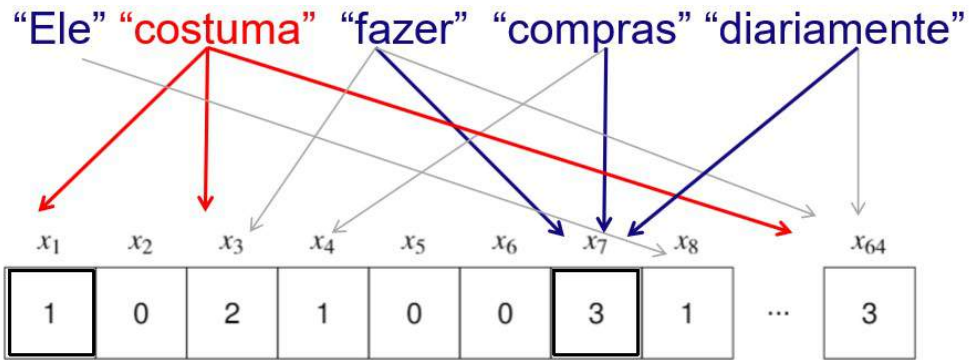


Figura II.6: Representação de uma sentença por meio de um vetor com as categorias do LIWC.

No LIWC é comum representar um texto ou uma sentença, em forma de um vetor cujo tamanho (*i.e.,* a quantidade de posições) corresponde ao seu total de categorias. Nesse caso, cada posição desse vetor contém um número inteiro representando a quantidade de palavras contabilizadas para cada categoria (*i.e.,* posição). A palavra “chorou”, por exemplo, se relaciona a cinco categorias de palavras: tristeza, emoção negativa, afeto, verbo e verbo no passado. Portanto, se a referida palavra for encontrada no texto analisado, cada uma dessas cinco categorias será incrementada em sua respectiva posição em um vetor [Pennebaker et al., 2015].

A Figura II.7 ilustra a representação vetorial de uma sentença *s* por meio de um vetor $\vec{v}$ de m posições, cada qual representando uma categoria do LIWC. Cada palavra *p* foi identificada em x_i e incrementada em sua posição (*i.e.,* categoria) correspondente no vetor $\vec{v}$. Nota-se que a palavra representada pela posição x_3 foi contabilizada 9 vezes e que a palavra representada pela posição x_m manifesta-se 7 vezes.

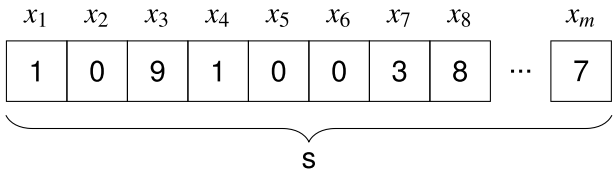


Figura II.7: Vetor $\vec{v}$ representando a frequência de palavras por categoria do LIWC em s.

A representação vetorial é de suma importância para a métrica proposta no presente trabalho, visto que cada sentença - seja ela candidata ou referência - precisa ser representada por um vetor de categorias com base no léxico em questão. Na literatura existem diversas medidas para calcular a similaridade (*i.e.,* compatibilidade) entre os dois vetores (*i.e.,* entre as duas sentenças), dentre as quais é possível destacar a similaridade do cosseno.

II.6 Similaridade do cosseno

Na Mineração de Textos (MT) há diversos estudos que utilizam a similaridade do cosseno para comparar textos ou sentenças e verificar sua similaridade, dentre os quais, é possível destacar [Di Thommazo et al., 2012, Santos et al., 2015]. A similaridade do cosseno é uma medida calculada a partir do produto interno entre dois vetores. Esses vetores representam textos ou sentenças e essa medida representa o ângulo entre os dois vetores (*i.e.,* duas sentenças ou dois textos). Logo, os valores possíveis para a similaridade do cosseno ficam entre 0 e 1, representando dissimilaridade e similaridade, respectivamente. O cálculo da similaridade do cosseno (SCos) se encontra descrito na Equação 9, em que *q* representa a quantidade de posições dos vetores *a* e *b* que devem ser do mesmo tamanho.

$$SCos = \frac{a_i \cdot b_i}{\sqrt{\sum_1^q a_i^2} \cdot \sqrt{\sum_1^q b_i^2}} \tag{9}$$

Conforme pode ser notado na equação acima, a similaridade do cosseno (SCos) é calculada em função da divisão do produtório das posições de *a* e *b* pelo produto do somatório das raízes de a_i e das raízes de b_i [de Oliveira et al., 2009].

II.7 Considerações

Neste capítulo, o objetivo foi esclarecer de que se tratam as Traduções Automáticas de Textos (TATs), em que áreas de estudos encontram-se situadas, bem como os desafios a serem superados por esse tipo de tradução. Em seguida foram apresentadas a maneira com que as TATs são avaliadas, algumas métricas existentes para realizar esse tipo de avaliação, suas limitações e as dificuldades enfrentadas por essas métricas, especialmente no que diz respeito à consideração de aspectos semânticos das traduções.

Ainda tratando das métricas existentes para avaliar TATs, foi dado um destaque ao estado da arte (*i.e.,* a métrica BLEU). Dado que essa métrica é comparada com a métrica proposta no presente trabalho, foi necessário apresentá-la em detalhes que vão desde sua formulação até o seu funcionamento, com ênfase aos *n-gramas* que são de fundamental importância para o entendimento do seu funcionamento. Por meio de exemplos detalhados foi possível ilustrar o funcionamento da BLEU além de suas limitações, que devem ser superadas pela métrica proposta neste trabalho.

Com o intuito de familiarizar o leitor com as técnicas utilizadas pela métrica proposta, foi apresentado o léxico afetivo do LIWC e a maneira com que representa vetorialmente uma sentença, identificando e contabilizando cada aspecto linguístico e psicológico presente em uma sentença. A ideia foi apresentar a possibilidade de adicionar semântica à avaliação de TATs. Depois de

apresentar a possibilidade de transformar sentenças em vetores que contabilizam, em cada uma de suas posições, aspectos linguísticos e psicológicos, foi apresentada a medida conhecida na literatura como similaridade do cosseno. Essa medida, muito utilizada para verificar a similaridade de textos e sentenças, foi apresentada com explicações sobre seu funcionamento, além de sua formulação.

Com a apresentação do LIWC pretendeu-se mostrar que existe uma alternativa para a adição de semântica às traduções. Em linhas gerais, este capítulo apresentou dificuldades encontradas pelas métricas atuais em considerar a semântica das traduções, com ênfase ao funcionamento da métrica BLEU, atual estado da arte. Por fim, com a similaridade do cosseno o objetivo foi mostrar que há na literatura, medidas de similaridade capazes de verificar a compatibilidade das sentenças transformadas em vetores de categorias psicolinguísticas gerados pelo LIWC e, consequentemente, apresentar indícios sobre a viabilidade da métrica proposta.

Capítulo III Trabalhos relacionados

Neste capítulo, serão descritos alguns trabalhos relacionados ao estudo que resultou nesta dissertação. Para tanto, são descritos os trabalhos que objetivam analisar a qualidade das TATs a partir de métricas e também um estudo que utiliza o LIWC para analisar formalidade e coesão em textos traduzidos do chinês para a língua inglesa.

III.1 Métricas utilizadas para avaliar traduções automáticas de texto

Há diversas métricas utilizadas para avaliar Traduções Automáticas de Textos TATs, algumas delas consideradas bem-sucedidas. Na busca por qualidade das TATs, os estudos que merecem maior destaque são efetuados pela *Defense Advanced Research Projects Agency* (DARPA). Nesses estudos, o objetivo foi formalizar métricas de adequação que se baseiam na correspondência da quantidade de informação transferida do texto original para o texto traduzido [White et al., 1994]. Em seu estudo, Linares [2005] cita algumas dessas métricas, dentre as quais destacou as métricas WER, PER, BLEU e *National Institute of Standards and Technology* (NIST), cujo acrônimo descreve o nome do Instituto que trabalhou para o seu desenvolvimento.

III.1.1 A métrica WER

A métrica *Word Error Rate* (WER), bem como a grande maioria das métricas, consiste em comparar a tradução de uma sentença referência (*i.e.,* tradução realizada por um humano especialista) com a tradução de uma sentença candidata (*i.e.,* tradução produzida por uma ferramenta de TAT). Trata-se de uma métrica bastante utilizada no reconhecimento de fala e é baseada na distância de Levenshtein [1966], porém, aplicada em palavras ao invés de letras. Sua fórmula é representada na Equação 10 a seguir.

$$WER = \frac{S+D+I}{N} \tag{10}$$

Os elementos presentes na Equação 10 da métrica WER são descritos da seguinte forma: `S` representa o total de `substituições`, `D` representa o total de `deleções`, `I` representa o total de `inserções` e `N` representa o total de `palavras contidas na tradução referência`.

O estudo de Beck [2009] indica que além disso, essa métrica penaliza bastante as trocas de

ordem, embora nem sempre isso represente necessariamente um indício de uma tradução ruim, especialmente em alguns idiomas em que a ordem de palavras é livre. O autor ainda exemplifica como essa penalização ocorre, em maiores detalhes. A figura III.1 exibe a aplicação da equação 10 em que S_{Ref} representa a sentença referência "O homem correu para dentro da casa dele" e S_{Cand} representa a sentença candidata "O homem correu para dentro de sua casa". Entre as referidas sentenças houve 3 `Substituições` (*i.e.,* "da" por "de", "casa" por "sua" e "dele" por "casa"), 0 `Deleções` e 0 `Inserções`.

$$WER(S_{Ref}, S_{Cand}) = \frac{3+0+0}{8} = \frac{3}{8} = 0.375$$

Figura III.1: Exemplo de aplicação da métrica WER.

Considerando que 0.0 representaria equivalência total e que 1.0 representaria uma incompatibilidade total entre as traduções, 0.375 reflete uma incompatibilidade consideravelmente alta para uma simples troca de ordem de duas palavras que não resultaram no comprometimento significativo das sentenças envolvidas. Vale ressaltar que a WER não considera `deleções` no caso de as sentenças referência e candidata serem do mesmo tamanho, como no exemplo utilizado.

A WER é uma métrica que favorece sentenças candidatas mais curtas. Isso ocorre por conta da divisão do número de edições pelo número de palavras na sentença referência, o que faz com que candidatas mais longas apresentem valores maiores do que 1 [Mauser et al., 2008]. Uma vez que essa métrica representa o nível de incompatibilidade entre as sentenças, quanto maior for esse valor, pior é a qualidade da tradução. Como os valores para a métrica WER precisam estar entre 0 e 1, nesses casos utiliza-se a Equação 10.1.

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+H} \tag{10.1}$$

Nessa nova equação, N é substituído por (S + D + H). Nesse caso, H significa um ajuste, ou seja, o número de `correções` entre a sentença referência e a sentença candidata. Tudo o que não for `substituição`, `deleção` ou `inserção` é contabilizado como `correção` (*i.e.* H) [Morris et al., 2004]. A figura III.2 exemplifica a identificação de cada item presente na Equação 10.1.

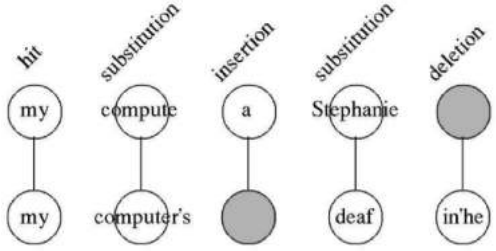


Figura III.2: Identificação de cada variável da WER. Fonte [Morris et al., 2004].

A figura III.3 apresenta a aplicação da Equação 10.1 ao exemplo ilustrado, onde H é representado por *"hit"*. Nesse exemplo é possível observar que a candidata menor é bem diferente da referência e que o cômputo do fator de `correção` H trouxe equilíbrio aos cálculos, mantendo a WER entre 0 e 1. Nesse caso específico o valor foi 1, representando incompatibilidade total.

$$WER(S_{Ref}, S_{Cand}) = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+H} = \frac{2+1+1}{2+1+1} = \frac{4}{4} = 1$$

Figura III.3: Exemplo de cálculo da métrica WER para candidatas menores.

III.1.2 A métrica PER

A métrica *Position-indepndent Error Rate* (PER) foi proposta por Tillmann et al. [1997] e trata-se de uma variação da métrica WER. Ao contrário da métrica WER, a métrica PER não considera a ordem das palavras. Sua fórmula, representada pela Equação 11, é a mesma da métrica WER para candidatas iguais ou maiores.

$$PER = \frac{S+D+I}{N} \tag{11}$$

Contudo, na PER, só se contabiliza `substituições` quando uma palavra é explicitamente substituída por outra, independente da ordem. Já em relação às `inserções` e `deleções`, elas só existem caso as sentenças referência e candidata sejam de tamanhos diferentes. Representando novamente o mesmo exemplo da seção III.1.1, em que a sentença referência é "O homem correu para dentro da casa dele" e a sentença candidata é "O homem correu para dentro de sua casa", percebe-se que houve apenas 2 `substituições` (*i.e.,* "da" por "de" e "dele" por "sua") ao invés de 3, uma vez que o substantivo "casa" não foi modificado. Também não houve `deleções` e nem `inserções`. A figura III.4 traz a aplicação da métrica PER para as referidas sentenças.

$$PER(S_{Ref}, S_{Cand}) = \frac{2+0+0}{8} = \frac{2}{8} = 0.25$$

Figura III.4: Exemplo de cálculo da métrica PER.

As métricas WER e PER tem relação direta com a distância entre os textos e seus valores e, como já foi dito, representam o grau de incompatibilidade entre as sentenças. Logo, quanto menor for o resultado dessas métricas, melhor é a qualidade da tradução avaliada.

III.1.3 A métrica NIST

A NIST surgiu a partir do estudo de Doddington [2002]. Trata-se de uma variação da métrica BLEU. Basicamente, a diferença é que a NIST considera a carga de informação contida em cada *n-grama* (*i.e.,* o quão "informativo" um *n-grama* é). Quanto menos frequente ele for em relação

ao texto, maior é o peso dado a ele. Logo, se um *n-grama* correto é identificado, quanto mais raro ele for, maior será sua importância e maior o seu peso (w), uma vez que ele representará um numerador maior na hora da divisão para o cálculo da precisão de *n-gramas*. A Equação 12 ilustra como esses pesos são calculados.

$$INFO(w_1...w_i) = \log_2 \left(\frac{\sum_{ocorrências\ de\ w_1...w_{i-1}}}{\sum_{ocorrências\ de\ w_1...w_i}} \right) \tag{12}$$

A outra diferença é que a NIST calcula a penalidade por brevidade (BP) minimizando seu impacto se a diferença entre as quantidades de termos da sentença referência e da sentença candidata for relativamente pequena [Santiago, 2013]. A Equação 13 mostra o cálculo da BP para a NIST.

$$BPNIST = e^{\left\{ FP \cdot \log_2 \left[min\left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}} \tag{13}$$

$FP = Fator\ de\ Penalidade = 0.5\ quando\ a\ ocorrência\ de\ palavras\ na\ candidata\ corresponder\ a\ \frac{2}{3}\ do\ número\ de\ palavras\ da\ referência\ ou\ menos.\ Caso\ contrário,\ FP = 1\ (i.e.,\ ausência\ de\ penalidade).$
$L_{ref} = quantidade\ de\ palavras\ no\ texto\ referência.$
$L_{sys} = quantidade\ de\ palavras\ no\ texto\ candidato.$

O cálculo completo da NIST pode ser visto na Equação 14.

$$NIST = \sum_{i=1}^{N} \left\{ \frac{\sum_{todas\ as\ co\ ocorrências\ de\ w_1...w_i} \cdot INFO(w_1...w_i)}{\sum_{todos\ os\ w_1...w_i\ no\ candidato}} \right\} \cdot BPNIST \tag{14}$$

$BPNIST = penalidade\ por\ brevidade\ do\ texto\ candidato.$
$N = 5.$

As métricas BLEU e NIST funcionam de forma inversamente proporcional às métricas WER e PER. No primeiro par, quanto maior for o valor obtido, maior é a compatibilidade entre as traduções. Já no segundo par, quanto maior for o valor obtido, maior é a incompatibilidade entre as traduções [Beck, 2009].

III.1.4 A métrica BLEU

A métrica BLEU é tida como o estado da arte com relação à avaliação de TATs. No entanto, de acordo com o estudo proposto em [de Melo et al., 2015], por ser baseada em pareamento exato e ordenado de palavras, 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. No mesmo estudo, a métrica BLEU foi usada 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. 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.

O estudo desenvolvido no presente trabalho se difere das métricas BLEU, NIST, WER e PER por introduzir a informação semântica das palavras. A nova métrica proposta (BRAPT) se assemelha à PER por não considerar a ordem das palavras, mas difere-se da mesma pela adição de semântica à avaliação. A 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 do léxico afetivo do LIWC. 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.2 Considerações

Neste capítulo, além da métrica BLEU abordada anteriormente, foram discutidas algumas métricas bastante utilizadas para avaliação de TATs. As avaliações realizadas por essas métricas acabam sempre se detendo a considerar tamanhos, medidas, quantidade e ordem dos termos. Como já foi mencionado nos capítulos anteriores, esses tipos de métricas não são suficientes para avaliar a qualidade de uma tradução, necessitando da intervenção humana que, nesse caso, pode representar uma pós-avaliação.

Outro aspecto a ser observado é que, embora existam ferramentas de análise textual como o LIWC, que possibilitam analisar aspectos linguísticos e psicológicos a partir de textos, ferramentas desse tipo não vem sendo utilizadas para avaliar a qualidade das TATs. Não foram encontrados estudos nessa direção. O objetivo principal do presente trabalho consiste em desenvolver uma nova métrica capaz de propor essa análise do ponto de vista semântico, até então desconsiderada pelas métricas existentes.

Capítulo IV Aspectos psicolinguísticos em traduções

Este capítulo consiste em apresentar uma funcionalidade do LIWC, além de uma contribuição relevante do presente estudo. Essa contribuição acrescenta uma abordagem individual acerca das categorias psicolinguísticas do LIWC. Este capítulo divide-se em duas seções. A seção IV.1 descreve uma funcionalidade da ferramenta LIWC que consiste na possibilidade de analisar cada aspecto linguístico e/ou psicológico presente em uma sentença. Dessa forma é possível verificar pontualmente a importância de cada aspecto linguístico ou psicológico de uma tradução referência e de uma tradução candidata, ou seja, é possível identificar o quão representativa uma determinada categoria (*i.e.,* determinado aspecto linguístico ou psicológico) é em relação à sentença estudada. A seção IV.2 apresenta uma contribuição relevante do presente trabalho. Essa contribuição consiste em um estudo mais aprofundado e individualizado (*i.e.,* categoria por categoria) sobre as mudanças ocorridas no processo de tradução, identificando especificamente os aspectos linguísticos e/ou psicológicos em que houve perdas ou acréscimos no referido processo.

IV.1 A representatividade de cada categoria em uma sentença

Com o intuito de analisar detalhadamente cada aspecto linguístico e psicológico que compõe uma determinada sentença (seja ela referência ou candidata), torna-se interessante verificar qual é o percentual de representatividade de cada categoria em relação a uma determinada sentença. Ao considerar a sentença referência (s_{Ref}) “Chove muito no momento e a população está apreensiva.”, antes mesmo de realizar uma análise minuciosa, nota-se a presença de algumas categorias: verbo, advérbio de intensidade, cognição, ansiedade, características relativas a tempo, artigo, preposição etc. No entanto, não é possível precisar o percentual de representatividade de cada categoria em relação à sentença. O percentual de representatividade da categoria cognição, por exemplo, é obtido considerando a quantidade de palavras presentes na sentença e que se enquadram nessa categoria e o somatório das palavras contabilizadas em todas as categorias. Algumas perguntas podem emergir:

- “Qual é o percentual de verbos?”
- “Qual é o percentual de cognição?”

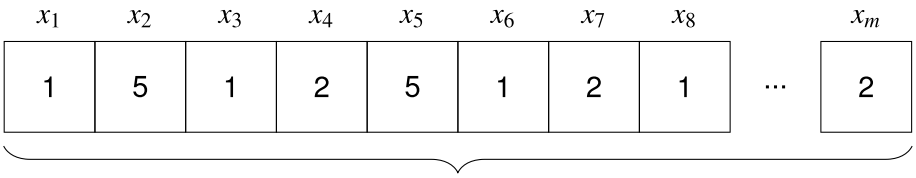
- “Qual é o percentual de emoções negativas?”

As respostas podem vir por meio da representação vetorial de cada uma dessas sentenças. A partir do léxico do LIWC, com a adição de uma categoria extra para palavras não contidas no referido léxico, é possível transformar uma sentença referência em um vetor de 65 posições, cada qual correspondente a uma das 64 categorias do LIWC ou a essa categoria extra. Logo, em cada posição desse vetor tem-se um valor inteiro que representa a quantidade de palavras contabilizadas para uma determinada categoria psicolinguística (*i.e.,* categoria que reflete um aspecto linguístico ou psicológico). O mesmo procedimento ocorre em relação à sentença candidata.

Dessa forma, torna-se possível analisar traduções operando com dois vetores de 65 posições (*i.e.,* um vetor referência e um vetor candidato), contendo valores inteiros guardados em cada uma de suas posições. Logo, cada posição desses vetores representa um aspecto linguístico ou psicológico da sentença a ser estudada e, consequentemente, a indicação do seu nível de importância na sentença.

IV.1.1 A frequência de palavras por categoria

A figura IV.1 apresenta a já citada sentença referência (s_{Ref}) representada agora por um vetor $\vec{v}_{Ref}$ em que $m = 65$. Nesse vetor, cada palavra foi contabilizada em diversas categorias. Ao somar os valores de todas as posições de $\vec{v}_{Ref}$, chega-se a um total de 38 contabilizações. A palavra “momento”, por exemplo, foi contabilizada em quatro categorias distintas: `palavras de função` (x_2), `aspectos relativos a espaço` (x_4), `relatividade` (x_5) e `aspectos relativos a tempo` (x_6). O procedimento é o mesmo tanto para uma sentença referência quanto para uma sentença candidata.



Vetor $\vec{v}_{Ref}$ com a frequência de palavras de s_{Ref} por categoria em que $m = 65$.

Figura IV.1: Representação vetorial da frequência de palavras por categoria na sentença s_{Ref}.

IV.1.2 O percentual de representatividade de cada categoria

A Equação 15 representa uma contribuição do LIWC. Com a aplicação dessa Equação em uma posição x_i de $\vec{v}_{Ref}$, é possível identificar o percentual de representatividade de determinada categoria em relação ao somatório de todas as posições do vetor $\vec{v}_{Ref}$. A partir da aplicação da referida Equação a cada uma das posições de $\vec{v}_{Ref}$ é possível produzir um segundo vetor

denominado $\vec{v}_{RefP}$. Esse novo vetor contém valores reais que correspondem ao percentual de representatividade de cada posição (*i.e.,* categoria) em relação à sentença s_{Ref}.

$$PRC(x_i) = 100 \cdot \left(\frac{x_i}{\sum_{i=1}^{m} x_i} \right) \tag{15}$$

A figura IV.2 ilustra o vetor $\vec{v}_{RefP}$ contendo os percentuais de representatividade de cada categoria, considerando que o somatório de todas as suas posições nesse exemplo é igual a 38. O somatório das posições visíveis na figura é igual a 20 e o somatório das posições não visíveis na figura (*i.e.,* das posições x_9 a x_{m-1}) é igual a 18. Logo, $18+20=38$.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8		x_m
2.63	13.15	2.63	5.26	13.16	2.63	5.26	2.63	···	5.26

Vetor $\vec{v}_{RefP}$ com os percentuais de representatividade de s_{Ref} por categoria.

Figura IV.2: Representação vetorial da sentença s_{Ref} em percentuais.

Considerando que a posição x_6 de $\vec{v}_{Ref}$ contém o valor 1 e que o somatório das posições do referido vetor é 38, o cálculo para obter o valor de x_6 em $\vec{v}_{RefP}$ pode ser visto na figura IV.3. Devido à multiplicação por 100, para o valor 0.0263, denota-se um percentual de 2.63%.

$$PRC(x_6) = 100. \left(\frac{1}{38} \right) = 2.63$$

Figura IV.3: Exemplo de cálculo do percentual de representatividade da categoria x_6 em $\vec{v}_{RefP}$.

Ao assumir que a posição x_5 de $\vec{v}_{RefP}$ representa a categoria `relativ` (relatividade), por exemplo, pode-se afirmar que nela cinco palavras foram contabilizadas e que 13.16% da sentença está representada pela categoria que reflete aspectos de relatividade. Sobre $\vec{v}_{RefP}$, é necessário salientar que trata-se de um vetor normalizado cujo somatório de suas posições é igual a 100. Retomando a sentença referência s_{Ref} “Chove muito no momento e a população está apreensiva”, cabe destacar, por meio da tabela IV.1, a categoria representada por cada posição do vetor $\vec{v}_{RefP}$ e seu percentual de representatividade em relação à referida sentença. Embora todas as operações e cálculos executados nesta seção façam menção a uma sentença referência, torna-se importante salientar que estas operações e cálculos seriam aplicados da mesma maneira em uma sentença candidata.

Tabela IV.1: Detalhamento das categorias representadas no vetor $\vec{v}_{RefP}$.

Posição	Categoria	Descrição	Percentual de representatividade
x_1	`124-humans`	Características humanas	2.63%
x_2	`1-funct`	Palavras de função	13.15%
x_3	`17-prep`	Preposições	2.63%
x_4	`252-space`	Espaço	5.26%
x_5	`250-relativ`	Relatividade	13.16%
x_6	`253-time`	tempo	2.63%
x_7	`150-ingest`	Ingestão	5.26%
x_8	`125-affect`	Afeto	2.63%
x_9 a x_{m-1}	`...`	Demais categorias	47.39%
x_m	`11-verb`	Verbos	5.26%

IV.1.3 Considerações sobre a representatividade de cada categoria

Esta seção apresenta a possibilidade de identificar o percentual de representatividade de cada categoria (*i.e.,* cada aspecto linguístico ou psicológico) contido em uma determinada sentença em que cada categoria está representada por uma posição de um vetor gerado pelo LIWC. Ao considerar a sentença referência “Ontem foi um dia maravilhoso, minha linda filha nasceu e eu estou muito feliz!”, é possível intuir que a mesma é composta por um grande percentual de emoções positivas. No entanto, qual seria esse percentual? Graças às ilustrações, cálculos e exemplos reais, essa pergunta foi respondida nesta seção.

Outra questão relevante a se considerar é que nem sempre os aspectos linguísticos e psicológicos podem estar expressos na sentença de maneira que se possa supor a dimensão de sua representatividade, como foi o caso do exemplo apresentado. Ao comparar essa sentença referência com uma eventual tradução candidata (*i.e.,* produzida por uma ferramenta de TAT), algumas perguntas podem emergir: quanto haveria de perda em relação às emoções positivas? Quanto haveria de perda ou ganho em relação a verbos, advérbios, afeto? Para que seja possível responder a essas perguntas também existe a necessidade de identificar as divergências ocorridas em cada categoria linguística ou psicológica (*i.e.,* divergências psicolinguísticas) presente nas sentenças durante o processo de tradução. Essas e outras perguntas podem ser respondidas com a proposição de uma solução capaz de suplantar essa necessidade. Essa solução é apresentada na seção seguinte.

IV.2 Cálculo de divergências psicolinguísticas por categoria

Esta seção apresenta uma das contribuições do presente trabalho que consiste na possibilidade de identificar o percentual de divergências psicolinguísticas verificado em cada categoria após o processo de tradução. Essa identificação ocorre por meio da comparação entre os dois vetores normalizados com base na Equação 15 apresentada na seção anterior. Esses vetores normalizados devem corresponder a uma sentença referência ($\vec{v}_{RefP}$) e a uma sentença candidata ($\vec{v}_{CandP}$). Para a identificação do percentual de divergências ocorrido no processo de tradução, foi desenvolvido o algorítimo calcDPC.

IV.2.1 O algoritmo calcDPC

O algoritmo `calcDPC` é capaz de gerar um vetor ($\vec{v}_{DivP}$) que calcula o percentual de divergências verificadas na comparação de cada posição (*i.e.,* cada categoria) dos vetores referência e candidato (*i.e.,* $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$). A Equação 16, apresentada a seguir, considera a posição i dos vetores referência e candidato (*i.e.,* $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$), para obter o percentual de divergências entre os valores contidos na posição i dos dois vetores normalizados. Esses vetores são representados por r e c na Equação.

$$Diverg(r_i, c_i) = 100\left(1 - \frac{min(r_i,\ c_i)}{max(r_i,\ c_i)}\right) \tag{16}$$

Logo, se temos o valor 15 (*i.e.,* 15%) na posição correspondente à categoria de cognição em $\vec{v}_{RefP}$ e o valor 3 (*i.e.,* 3%) na mesma posição em $\vec{v}_{CandP}$, isso significa que em relação à categoria que representa cognição, o percentual de divergências entre os dois vetores (*i.e.,* as duas sentenças) é de 80%. Supondo que os referidos vetores são representados por r e c e que a categoria em questão é representada pela posição 9 de ambos os vetores, a aplicação da Equação 16 é exemplificada na figura IV.4.

$$Diverg(r_9, c_9) = 100\left(1 - \frac{3}{15}\right) = 80.0$$

Figura IV.4: Identificação de divergências na posição 9 dos vetores $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$.

Como o valor contido na referida categoria de $\vec{v}_{RefP}$ é maior do que o valor contido na mesma categoria de $\vec{v}_{candP}$, é possível dizer também, que houve uma perda de 80% de aspectos cognitivos na referida tradução. O algoritmo 1 exibe a representação dos passos executados pelo `calcDPC`.

O `calcDPC` recebe como argumentos dois vetores normalizados (*i.e.,* $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$), correspondentes às sentenças referência e candidata, possibilitando identificar aspectos semânticos (*i.e.,* aspectos linguísticos e psicológicos) contidos nas duas sentenças. Em seguida, executa-se os seguintes passos:

Algorithm 1 calcDPC($\vec{v}_{RefP}$, $\vec{v}_{candP}$)

Input:

- $\vec{v}_{RefP}$ = Vetor normalizado representando a sentença referência
- $\vec{v}_{CandP}$ = Vetor normalizado representando a sentença candidata

Output: $\vec{v}_{DivP}$, vetor com as divergências verificadas em cada categoria de $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$

1: $t \leftarrow \text{obtemT}(\vec{v}_{RefP})$
2: $\vec{v}_{DivP} \leftarrow \text{criaVetorDivergencias}(t)$
3: **for** ($i = 1$ to t) **do**
4: $\quad \vec{v}_{DivP}[i] \leftarrow \text{divergenc}(\vec{v}_{RefP}[i], \vec{v}_{CandP}[i])$
5: **end for**
6: **return** $\vec{v}_{DivP}$

- Passo 1: A partir da função `obtemT`, a variável t recebe o tamanho do vetor $\vec{v}_{RefP}$, que é o mesmo de $\vec{v}_{CandP}$, visto que as sentenças são representadas por vetores de mesmo tamanho;

- Passo 2: A partir da função `criaVetorDivergencias`, que recebe t como argumento, obtem-se o vetor de divergências $\vec{v}_{DivP}$ com t posições inicializadas com o valor 0.0;

- Passo 3: Inicia-se um laço de repetição de 1 até t, permitindo percorrer cada posição dos vetores $\vec{v}_{DivP}$, $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$;

- Passo 4: A função `divergenc` recebe como argumento os valores contidos na posição i de $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$. Em seguida, com o auxílio da Equação 16, a referida função retorna para a posição i de $\vec{v}_{DivP}$ o percentual de divergências verificado entre os dois valores passados como argumento;

- Passo 5: Fim do laço de repetição;

- Passo 6: O algoritmo `calcDPC` retorna o vetor $\vec{v}_{DivP}$ preenchido com o percentual de divergências verificado em cada posição (*i.e.,* categoria) dos vetores $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$.

IV.2.2 Exemplo das divergências identificadas por categoria

Os vetores $\vec{v}_{Ref}$ de uma sentença referência “Chove muito **no momento** e a população está **apreensiva**.” e $\vec{v}_{Cand}$ de uma sentença candidata “Chove muito e a população está **preocupada**.”, tiveram, respectivamente, 38 e 30 palavras contabilizadas em suas posições. A figura IV.5 contém os vetores normalizados $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$ das referidas sentenças, além do vetor $\vec{v}_{DivP}$, que representa as divergências verificadas em cada posição (*i.e.,* categoria) desses vetores. A referida figura também ilustra o vetor $\vec{v}_{DivP}$ retornado pelo algoritmo `calcDPC`. Nesse vetor é possível observar que houve divergências verificadas em todas as categorias.

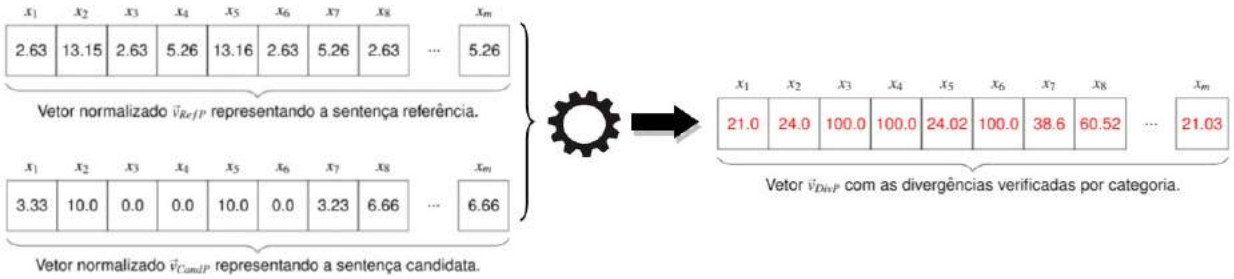


Figura IV.5: Divergências psicolinguísticas identificadas por categoria.

Cabe lembrar que os três vetores possuem 65 categorias (*i.e.,* posições) cada um, embora nem todas estejam presentes na ilustração. Utilizando a posição x_5 (*i.e.,* categoria que reflete aspectos de relatividade) dos vetores $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$, a aplicação da Equação 14 pode ser vista novamente na figura IV.6.

$$Diverg(r_5, c_5) = 100\left(1 - \frac{10.0}{13.16}\right) = 24.02$$

Figura IV.6: Verificação de divergências na posição 5 $\vec{v}_{RefP}$ (r) e $\vec{v}_{CandP}$ (c).

Novamente, como o maior valor (13.16) estava na sentença referência e o menor valor (10.0) estava na sentença candidata, pode-se afirmar, também, que no processo de tradução automática houve uma perda de 24.02% em relação à categoria que reflete aspectos de relatividade.

A tabela IV.2 ilustra as seguintes informações acerca das sentenças em questão: a descrição da categoria representada por cada posição x_i, a frequência de palavras em $\vec{v}_{Ref}$ e $\vec{v}_{Cand}$, o percentual de representatividade de cada categoria (*i.e.,* cada posição x_i) em $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$ e, em vermelho, o vetor $\vec{v}_{DivP}$ contendo as divergências apuradas entre cada posição x_i de $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$.

Tabela IV.2: Detalhamento dos vetores $\vec{v}_{Ref}$, $\vec{v}_{Cand}$, $\vec{v}_{RefP}$, $\vec{v}_{CandP}$ e $\vec{v}_{DivP}$.

Posição	**Categoria**	$\vec{v}_{Ref}$	$\vec{v}_{Cand}$	$\vec{v}_{RefP}$	$\vec{v}_{CandP}$	$\vec{v}_{DivP}$
x_1	`124-humans`	1	1	2.69%	3.33%	+21.0%
x_2	`1-funct`	5	3	13.15%	10.0%	−24.0%
x_3	`17-prep`	1	0	2.63%	0.0%	−100.0%
x_4	`252-space`	2	0	5.26%	0.0%	−100.0%
x_5	`250-relativ`	5	3	13.16%	10.0%	−24.02%
x_6	`253-time`	1	0	2.63%	0.0%	−100.0%
x_7	`150-ingest`	2	1	5.26%	3.23%	−38.6%
x_8	`125-affect`	1	2	2.63%	6.66%	+60.52%
x_9 a x_{m-1}	`...`	–	–	–	–	–
x_m	`11-verb`	2	2	5.26%	6.66%	+21.03%

Ao observar a tabela IV.2, é possível notar que o vetor $\vec{v}_{DivP}$ apresentou divergências em todas as categorias ilustradas. Ao observar especificamente a categoria representada pela posição x_m (`11-verb`), percebe-se que apesar dessa categoria ter contabilizado 2 palavras tanto no vetor re-

ferência ($\vec{v}_{Ref}$) quanto no vetor candidato ($\vec{v}_{Cand}$), houve um percentual de divergências de 21.03% computado em $\vec{v}_{DivP}$. Isso ocorre por que o vetor referência tem mais palavras contabilizadas (38) do que o vetor candidato (30). Logo, a referida categoria é mais representativa com 6.66% no vetor candidato $\vec{v}_{CandP}$ do que com 5.26% no vetor referência $\vec{v}_{RefP}$. Lembrando que $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$ são os vetores considerados para o cálculo das divergências em $\vec{v}_{DivP}$.

O vetor $\vec{v}_{DivP}$ sempre vai conter, em cada uma de suas posições, valores entre 0.0 e 100.0. O valor 0.0 indica que não houve divergências psicolinguísticas na mesma posição (*i.e.,* mesma categoria) entre as duas sentenças e o valor 100.0 indica que houve 100% de divergências verificadas na mesma posição (*i.e.,* categoria) de ambas as sentenças.

IV.2.3 Considerações sobre as divergências identificadas

Neste capítulo, na seção IV.1 foi possível abordar a importância de identificar os aspectos psicolinguísticos presentes em uma sentença e, principalmente, poder apresentar por meio de um vetor normalizado, o percentual de representatividade (*i.e.,* o grau de importância) de cada um desses aspectos para a referida sentença. Já na seção IV.2, o algorítimo `calcDPC` mostrou-se capaz analisar os vetores normalizados gerados na seção anterior e produzir um novo vetor, contendo, em cada uma de suas posições (*i.e.,* em cada uma de suas categorias psicolinguísticas), as divergências verificadas entre uma sentença referência e uma sentença candidata. Por meio desse vetor, pode ser verificado o percentual de divergências ocorrido em cada aspecto linguístico ou psicológico durante o processo de tradução.

Se no processo de tradução, a categoria que reflete emoções negativas aparece com um percentual de representatividade menor na sentença candidata (representada por $\vec{v}_{CandP}$) quando comparada com a sentença referência (representada por $\vec{v}_{RefP}$), isso significa que houve uma perda percentual em relação às emoções negativas. Se a sentença candidata apresentar um percentual de representatividade maior na referida categoria, é possível afirmar que houve um acréscimo em relação às emoções negativas no referido processo. Em ambas as situações é possível afirmar que houve divergências e, principalmente, é possível quantificar percentualmente essas divergências. O mesmo vale para as demais categorias (*e.g.,* verbos, preposições, pronomes, afeto, ansiedade, raiva).

Ainda tratando dos valores contidos em determinada categoria dos vetores normalizados $\vec{v}_{RefP}$ e $\vec{v}_{CandP}$, se ambos forem iguais, é possível afirmar que não houve mudanças em relação a determinada categoria. Caso uma categoria apresente alguma representatividade na sentença referência e não apresente representatividade alguma na sentença candidata, é possível afirmar que houve uma perda total de determinado aspecto linguístico ou psicológico, como aconteceu, por exemplo, com a categoria que reflete aspectos relativos a tempo, ilustrado na figura IV.5 e,

posteriormente na tabela IV.2. Nesse momento, uma diferença de 15.0% para 0.0% ou de 50.0% para 0.0% representam igualmente uma perda de 100% de determinado aspecto psicolinguístico ocorrida no processo de tradução. Nessas duas situações trata-se apenas de uma análise individual de cada categoria, não significando, portanto, que ambas terão o mesmo peso ao realizar o cômputo da compatibilidade entre as duas sentenças na métrica proposta.

As métricas existentes para realizar avaliação de Traduções Automáticas de Texto (TATs) fornecem apenas um valor correspondente ao nível de compatibilidade entre uma sentença referência e uma sentença candidata. O propósito dessas métricas, incluindo a métrica proposta no presente estudo, é o de informar sobre a compatibilidade entre as duas sentenças, sem no entanto oferecer a possibilidade de uma análise mais profunda sobre as características dessas sentenças ou até mesmo uma análise mais pontual, acerca de um determinado aspecto linguístico ou psicológico presente nas traduções. Essa lacuna é preenchida pelo algoritmo `calcDPC`.

É importante destacar que não é função do algoritmo `calcDPC` verificar a compatibilidade entre sentenças, mas sim realizar uma análise individual de cada categoria (*i.e.,* de cada aspecto psicolinguístico). Esse algoritmo tem a função de identificar e possibilitar uma análise minuciosa acerca das divergências ocorridas no processo de tradução, tanto das divergências apresentadas em aspectos linguísticos (*e.g.,* pronomes, advérbios, substantivos, preposições) quanto em aspectos psicológicos (*e.g.,* emoções positivas, raiva, ansiedade, aspectos cognitivos). Com isso, pode ser possível avaliar até mesmo as diferenças (*i.e.,* divergências) linguísticas e psicológicas eventualmente geradas por determinadas ferramentas de TAT.

Essa possibilidade pode significar uma contribuição relevante do referido algoritmo, sinalizando abordagens que podem ser muito úteis para investigar que tipo de problemas estão ocorrendo no processo de tradução de cada uma dessas ferramentas e talvez, quem sabe, até auxiliar a melhorá-las. O próximo passo é realizar experimentos com textos traduzidos por ferramentas de TAT objetivando confirmar a eficácia da contribuição apresentada neste capítulo.

Capítulo V A métrica BRAPT

Considerar aspectos inerentes à capacidade humana de interpretação e avaliação são necessidades reais que podem melhorar a qualidade das métricas utilizadas para avaliar TATs. Aspectos linguísticos e psicológicos precisam estar mais presentes nos processos de avaliação da qualidade desse tipo de tradução [Sales, 2011]. Este capítulo descreve a métrica *Bilingual Rating of Psycholinguistic Perspectives in Translations* (BRAPT), principal contribuição do presente trabalho.

V.1 Metodologia

A métrica BRAPT pode ser utilizada tanto para Traduções Diretas (TD) quanto para Traduções Inversas (TI). No entanto, no presente trabalho os estudos e experimentos estão concentrados nas traduções da língua inglesa para o português do Brasil (*i.e.,* Tradução Direta (TD)). Assim como a grande maioria das métricas citadas, a métrica BRAPT consiste em comparar uma tradução referência (tida como confiável) com uma tradução candidata cuja qualidade se pretende avaliar. O diferencial é que a métrica BRAPT tem como ponto central a observação de aspectos linguísticos e psicológicos presentes nas traduções. Esses aspectos são considerados por meio de um léxico afetivo capaz de classificar palavras de acordo com categorias que refletem esses aspectos linguísticos ou psicológicos, adicionando, dessa forma, semântica ao processo de avaliação. No caso específico deste estudo, o léxico adotado é proveniente da ferramenta LIWC em sua versão de 2007 para o PB.

O LIWC em PB de 2007 possui 64 categorias . Além de utilizar todas essas categorias, optou-se por acrescentar uma categoria extra para contabilizar palavras não existentes na referida ferramenta. Apesar de contar com mais de 127.000 palavras na versão para o português do Brasil, foi constatado que palavras relevantes não foram encontradas na referida versão desse léxico. Ainda precisam ser feitos experimentos para verificar se a ausência de palavras ocorre com a mesma frequência e relevância na versão do LIWC para a língua inglesa. Com o auxílio do LIWC, sentenças referência e candidata são transformadas em vetores de frequência de palavras por categorias. Essas categorias refletem aspectos linguísticos e psicológicos contidos nas sentenças, adicionando semântica à avaliação.

Como já foi mencionado, com o auxílio da ferramenta LIWC, na referida métrica, textos - sejam eles referência ou candidato - precisam ser separados em sentenças e estas, por sua vez, precisam ser transformadas em vetores de 65 posições em que cada uma dessas posições representando um aspecto linguístico ou psicológico. Logo, a BRAPT opera com 2 vetores de 65 posições representando uma sentença referência tida como confiável ($\vec{v}_{Ref}$) e uma sentença candidata ($\vec{v}_{cand}$), geralmente produzida por uma ferramenta de TAT, e cuja qualidade se pretende avaliar.

V.1.1 O algoritmo calcBRAPT

O algoritmo 2 apresenta, em detalhes, os passos executados pelo `calcBRAPT`. Esse algoritmo recebe como argumentos as sentenças referência (s_{ref}) e candidata (s_{cand}), além do léxico (lex) a ser utilizado para realizar a identificação dos aspectos semânticos (*i.e.,* dos aspectos linguísticos e psicológicos) contidos nas referidas sentenças. Depois de executar os passos de 1 a 4, o referido algoritmo retorna a compatibilidade BRAPT com o auxílio da medida conhecida como similaridade do cosseno.

Algorithm 2 calcBRAPT(s_{ref}, s_{cand}, lex)

Input:

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

Output: compat, a compatibilidade entre s_{ref} e s_{cand}

1: $\text{compat} \leftarrow \emptyset$
2: $\vec{v}_{ref} \leftarrow \text{termosPorCat}(s_{ref}, lex)$
3: $\vec{v}_{cand} \leftarrow \text{termosPorCat}(s_{cand}, lex)$
4: $\text{compat} \leftarrow \text{simCos}(\vec{v}_{ref}, \vec{v}_{cand})$
5: **return** compat

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, considerando que m representa o total de categorias presentes em lex. 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 na geração de $\vec{v}_{cand}$ com base em s_{cand}. Dessa forma, a partir desse momento, os vetores $\vec{v}_{Ref}$ e $\vec{v}_{Cand}$ passam a conter, respectivamente, a frequência de palavras por categoria de s_{ref} e s_{cand}. Os passos 4 e 5 representam a verificação e o retorno da compatibilidade (*i.e.,*

similaridade) entre os dois vetores com o auxílio da função simCos, que retorna a similaridade do cosseno entre $\vec{v}_{ref}$ e $\vec{v}_{cand}$.

A figura V.1 exemplifica o cômputo da compatibilidade (similaridade) BRAPT entre dois vetores (*i.e.,* duas sentenças). Os vetores hipotéticos x e y poderiam ser referentes, respectivamente, às sentenças referência e candidata.

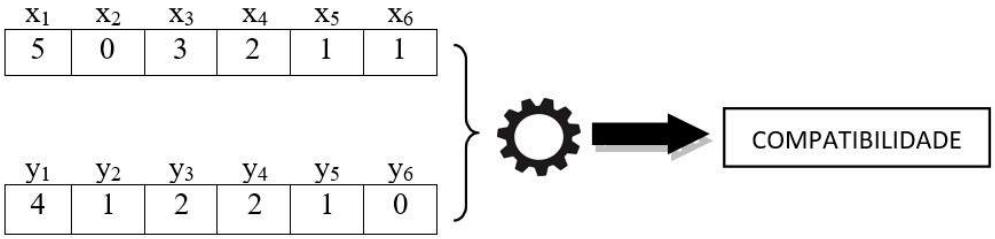


Figura V.1: Compatibilidade BRAPT entre dois vetores de seis posições.

O algoritmo `calcBRAPT` opera em função do número de categorias do léxico utilizado, que nesse caso hipotético é cinco (*i.e.,* $m = 5$). Com a adição da categoria extra para contabilizar as palavras não existentes no léxico, o algoritmo `calcBRAPT` trabalha com vetores de seis posições (*i.e.,* $m = (m+1) = 6$).

Relembrando a Equação (9) da seção II.6, a compatibilidade (*i.e.,* similaridade do cosseno) verificada entre os vetores hipotéticos x e y é representada na figura V.2 por $SCos$. Nessa figura é possível verificar o desenvolvimento detalhado dos cálculos resultantes da aplicação dessa medida de similaridade, cujo resultado foi de $0,96$. Logo, a compatibilidade BRAPT verificada foi de 0.96.

$$SimCos(x,y) = \frac{x_i \cdot y_i}{\sqrt{\sum_1^q x_i^2} X \sqrt{\sum_1^q y_i^2}}$$

$$SimCos(x,y) = \frac{(5X4)+(0X1)+(3X2)+(2X2)+(1X1)+(1X0)}{\sqrt{5^2+0^2+3^2+2^2+1^2+1^2} X \sqrt{4^2+1^2+2^2+2^2+1^2+0^2}}$$

$$SimCos(x,y) = \frac{20+0+6+4+1+0}{\sqrt{40X\sqrt{26}}}$$

$$SimCos(x,y) = \frac{31}{6,33X5,10}$$

$$SimCos(x,y) = \frac{31}{32,29} = 0,96$$

Figura V.2: Verificação da similaridade do cosseno entre dois vetores (x e y) de seis posições.

V.2 Considerações

Neste capítulo foi possível descrever o funcionamento da métrica BRAPT, bem como a maneira com a qual a mesma é capaz de adicionar semântica à avaliação de TATs, com o auxílio do léxico do LIWC em PB. Anteriormente, também foi possível verificar como esse léxico se comporta e a maneira com que transforma sentenças - sejam elas referência ou candidata - em vetores psicolinguísticos de frequência de palavras por categoria. A partir desses vetores, as palavras de determinada sentença são contabilizadas em categorias que refletem aspectos linguísticos e psicológicos, tornando possível, a partir desse ponto, adicionar semântica às sentenças (*i.e.,* vetores) a serem avaliadas.

A verificação da compatibilidade entre uma sentença referência e uma sentença candidata é apresentada em detalhes pelo algoritmo `calcBRAPT` ao longo de cada um de seus cinco passos. Por fim, com o auxílio de um exemplo hipotético dos vetores x e y representando, respectivamente, uma sentença referência e uma sentença candidata (ilustrado pela figura V.1), foi possível simular a aplicação da métrica BRAPT, apresentando, inclusive, o cálculo detalhado da compatibilidade entre x e y, com a aplicação da similaridade do cosseno. Dessa forma, neste capítulo é possível conhecer a métrica proposta no presente estudo e perceber a maneira simples e objetiva com que a mesma funciona e atinge seu propósito, que é o de acrescentar semântica à avaliação de TATs.

Resta, no entanto, estabelecer comparações entre a métrica BRAPT e a métrica BLEU, atual estado da arte, a fim de colocar à prova sua eficácia, especialmente em situações em que as métricas atuais se mostram deficientes. Essas comparações precisam ser realizadas com textos traduzidos por humanos (especialistas em tradução) e com textos produzidos por ferramentas de TAT conhecidas e amplamente utilizadas em ambiente *web*. A fim de validar as comparações e atestar a consistência dos resultados, faz-se necessário que os testes sejam supervisionados por um profissional especializado em traduções.

Capítulo VI Experimentos

Neste capítulo são apresentados os experimentos realizados com o propósito de avaliar as contribuições do presente trabalho. Inicialmente é realizada uma investigação acerca da representatividade de cada categoria psicolinguística em relação aos 10 textos utilizados neste estudo. Esses 10 textos possuem 128 sentenças (*i.e.,* uma média de 12.8 sentenças por texto). A seção VI.5 traz uma comparação entre o atual estado da arte (*i.e.,* a métrica BLEU) e a métrica BRAPT, proposta no presente estudo. Essas comparações são realizadas sob a supervisão de um especialista em traduções.

As métricas para avaliação de TATs são de fundamental importância e merecem um estudo aprofundado acerca de seu desempenho e de eventuais melhorias na forma como verificam a compatibilidade entre sentenças referência e candidata. No entanto, ao verificar a compatibilidade, tais métricas atuam apenas no sentido de identificar problemas ocasionados por traduções ruins. Muitas vezes, como já foi mencionado, essas métricas falham por não serem capazes de identificar determinados aspectos como a semântica das sentenças avaliadas, por exemplo. O objetivo principal deste capítulo é justamente investigar as razões que fazem com que as ferramentas de TAT ainda produzam traduções distantes de traduções realizadas por especialistas humanos e, consequentemente, identificar problemas que fazem com que as métricas existentes falhem na tarefa de avaliar esse tipo de tradução de forma eficaz.

É possível chegar a algumas conclusões com base nos textos avaliados neste capítulo e a partir de algumas perguntas que poderão nortear as investigações: quais são os aspectos que fazem com que ferramentas de TAT produzam traduções com problemas? Quais são os aspectos que não estão sendo observados por métricas que avaliam TATs? Que tipo de diferenças estão ocorrendo no processo de tradução?

Para tentar responder a algumas dessas perguntas são investigados, na seção VI.1, os aspectos psicológicos e linguísticos mais representativos em traduções produzidas por humanos especialistas. Na seção VI.2, são identificadas as perdas mais significativas de determinados aspectos psicológicos nas traduções produzidas por cada ferramenta de TAT. A seção VI.3, traz as divergências apresentadas por cada ferramenta de TAT acerca de aspectos linguísticos presentes no LIWC. Essas análises devem ajudar a identificar os fenômenos ocorridos no processo de tradução automática de textos. Aspectos esses que precisam ser identificados pelas métricas

e que podem estar influenciando negativamente na qualidade das TATs. Por fim, na seção VI.4 são feitas as considerações acerca das investigações e experimentos realizados neste capítulo, além dos passos seguintes a serem tomados.

VI.1 Aspectos linguísticos e psicológicos mais representativos

Esta seção tem o objetivo de evidenciar os aspectos linguísticos e psicológicos mais representativos identificados nos dez textos utilizados no presente estudo. Durante os experimentos foi possível identificar que, assim como ocorre com os aspectos linguísticos, os aspectos psicológicos também estão muito presentes nas traduções, bem como as perdas mais acentuadas de aspectos psicológicos e linguísticos nos referidos textos. Nesse sentido, ao realizar os testes com todas as sentenças traduzidas por um especialista humano foi possível identificar o percentual médio de representatividade dos aspectos linguísticos e psicológicos encontrados nas referidas sentenças.

A seção VIII.1 do anexo VIII traz os percentuais de representatividade de todas categorias do LIWC em relação a esses textos. A figura VI.1 apresenta o gráfico com os percentuais de representatividade dos dez aspectos psicolinguísticos (*i.e.,* linguísticos e psicológico)s mais relevantes encontrados nas traduções referência realizadas por um especialista, dentre os quais é possível destacar `1-funct`, `131-cogmech` e `250-relativ`.

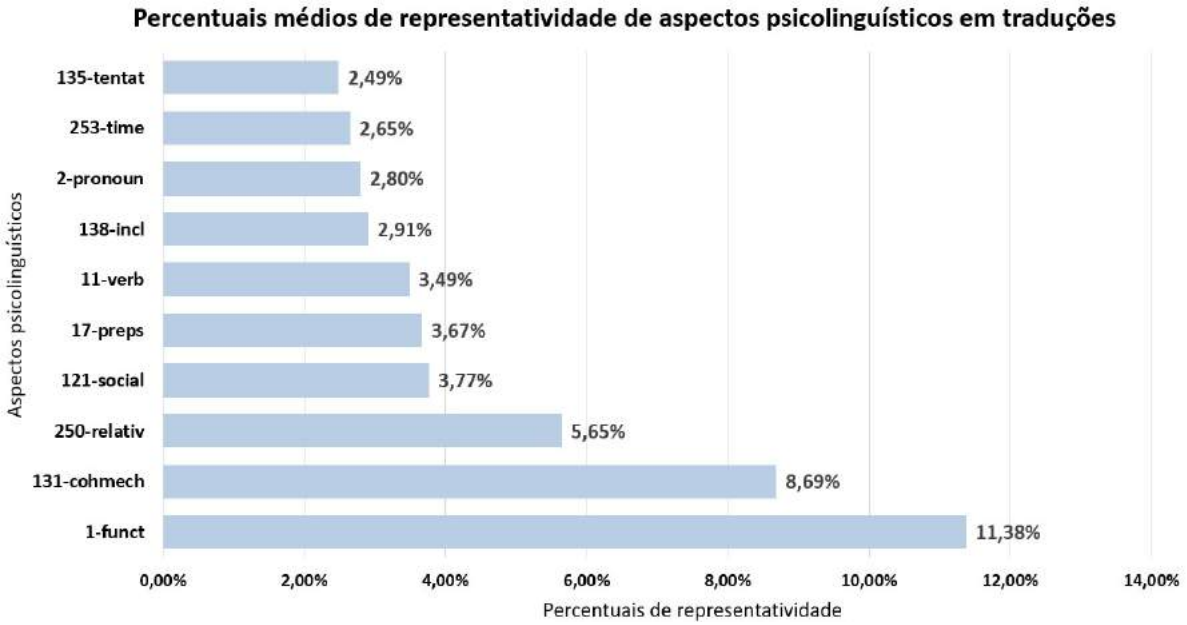


Figura VI.1: Representatividade de aspectos psicolinguísticos em traduções.

Os resultados mostraram-se consistentes com o estudo de Miller [1991], que apontam as palavras de função (*i.e.,* `1-funct`) como as mais representativas em qualquer tipo de comunicação, seja ela escrita ou falada. Além das palavras de função, com 11.38%, também cabe destacar a

representatividade das categorias `131-cogmech` (*i.e.,* cognição) com 8.69% e `250-relativ` (*i.e.,* relatividade) com 5.65%.

Uma vez que foram identificados os percentuais médios de representatividade das principais categorias do LIWC nas referidas traduções, o próximo passo é verificar se estão ocorrendo perdas em relação a esses e outros aspectos psicolinguísticos após o processo de tradução de sentenças referência por cada uma das ferramentas de TAT estudadas no presente trabalho.

VI.2 Análise de aspectos psicológicos

Conforme apresentado na seção anterior, os aspectos psicológicos também são muito presentes em traduções. Dessa forma, as investigações conduzidas nesta seção consistem em identificar, com o auxílio do algoritmo `calcDPC`, apresentado na seção IV.2.2, os dez aspectos psicológicos com maiores percentuais médios de perdas identificados nas traduções produzidas pelas ferramentas BI, GT e WL. Na seção VIII.2 do capítulo VIII é possível encontrar a lista completa com as divergências verificadas em cada categoria psicolinguística para as três ferramentas de TAT estudadas. A figura VI.2 apresenta o gráfico com os resultados dessa investigação.

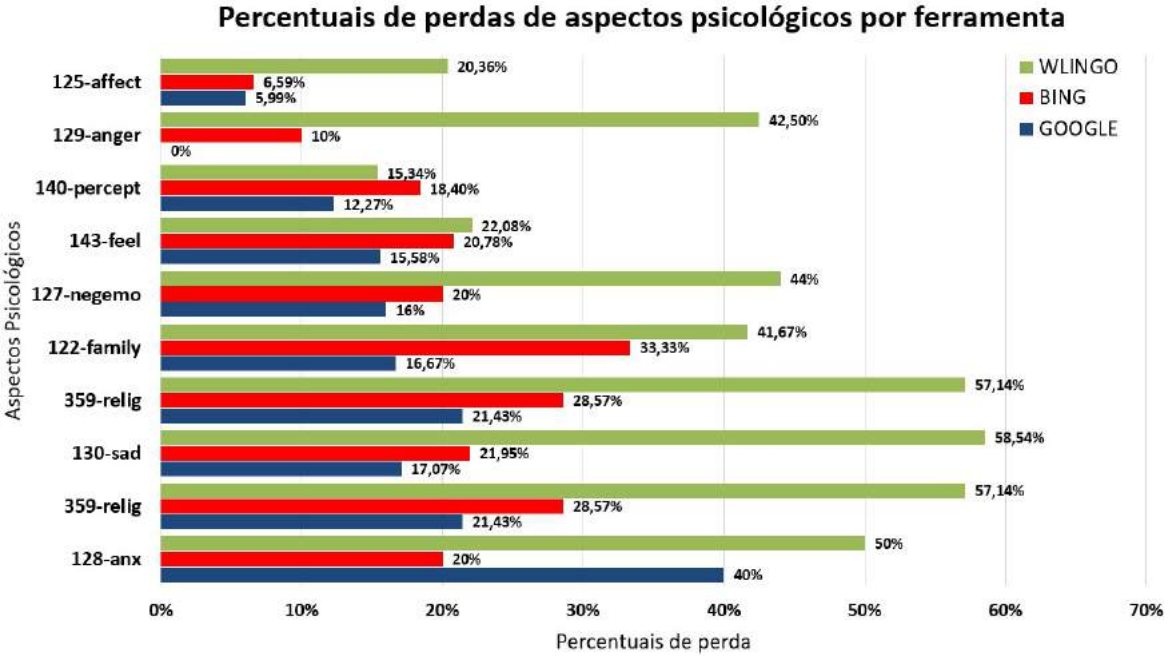


Figura VI.2: Perdas de aspectos psicológicos em ferramentas de TAT.

Com exceção das categorias `143-feel` (sentimento) e `140-percept` (percepção), no gráfico é possível observar que a ferramenta WL (já identificada como a menos sofisticada dentre as três ferramentas estudadas) apresentou percentuais médios de perdas bem superiores aos apresentados pelas ferramentas BI e GT em todos os demais aspectos.

No mesmo gráfico também é possível notar que o GT apresentou uma média de **40.0%** de perdas contra 20.0% do BI em para a categoria `128-anx` (ansiedade). Essas perdas, no entanto,

ocorreram em apenas sete sentenças em que uma palavra foi contabilizada tanto na sentença referência quanto na candidata produzida pelo GT. As diferenças ficaram por conta da representatividade. Nas sentenças produzidas pelo GT essa categoria se mostrou menos representativa. Na sentença 2 do texto 6, por exemplo, a categoria `128-anx` apresentou 0.69% de representatividade na sentença produzida pelo GT contra 0.89% na sentença referência (*i.e.,* 22.47% de perda).

Também vale destacar, de forma positiva, o fato da ferramenta GT não ter apresentado perdas para a categoria `129-anger` (raiva), enquanto as ferramentas BI e WL apresentaram uma média de perdas de 10.0% e 42.5%, respectivamente. A tabela VI.1 traz um exemplo das perdas com a ferramenta WL na sentença 8 do texto 10. A categoria `129-anger` (raiva), nesse exemplo, se associa às palavras "ameaças" e "cortar".

As seguintes formatações são utilizadas para a referida tabela e para as tabelas subsequentes em que determinada categoria é analisada em relação a uma ferramenta de TAT: As palavras pertencentes à categoria analisada estão sublinhadas. As divergências entre as sentenças referência e candidata estão destacadas em **negrito**. As colunas **"fp"** e **"rep"** representam, respectivamente, a frequência de palavras e a representatividade da referida categoria para cada sentença analisada.

Tabela VI.1: Perdas na categoria `raiva` com o WL (Texto 10, sentença 8).

Sentença	**Perda de 44.87% na categoria `129-anger` com o WL**	**fp**	**rep**
Refer.	**As** técnicas para **motivar os jogadores incluíam** ameaças **de cortar suas pernas e jogá-las a** cães **famintos.**	2	2.63%
Cand. WL	Técnicas para **motivate** ameaças **incluídas jogadores para eliminar seus pés e para jogá-los para** cães **ravenous.**	1	1.45%

Em relação à ferramenta BI, na categoria `12-anger` houve casos em que a perda dessa característica em determinadas sentenças chegou ao percentual de 38.96%. Já na categoria `122-family` (aspectos relacionados à família) o BI apresentou média de **33.33%** de perdas contra 16.67% do GT. Houve casos em que a perda de aspectos familiares em uma sentença produzida pela BI chegou ao percentual de 44.86%

Ao abordar a ferramenta WL, cabe destacar o alto percentual médio de perdas (*i.e.,* **58.54%**) apresentado por essa ferramenta para a categoria `130-sad` (tristeza). A tabela VI.2 apresenta uma sentença em que essa perda chegou a 100.0% na referida categoria para determinada sentença.

Tabela VI.2: Perdas na categoria `tristeza` com o WL (Texto 1, sentença 7).

Sentença	**Perda de 100.0% na categoria `130-sad` com o WL**	**fp**	**rep**
Refer.	**A verdadeira depressão** é diferente **da sensação de tristeza** em **muitos sentidos.**	2	3.13%
Cand. WL	**O depression real** é diferente **dos azuis** em **diversas maneiras chaves.**	0	0.0%

VI.3 Análise de aspectos linguísticos

Com o intuito de concluir as investigações do presente trabalho, nesta seção buscou-se identificar, também por meio do algoritmo `calcDPC`, a média de divergências ocorridas em aspectos linguísticos presentes nas referidas traduções. Vale lembrar, novamente, que a seção VIII.2 do capítulo VIII apresenta uma lista completa com as divergências verificadas em cada categoria psicolinguística para as três ferramentas de TAT estudadas. A tabela VI.3 apresenta, para cada uma das três ferramentas de TAT, as médias de divergências (perda ou acréscimo) observadas em cada categoria linguística (*i.e.,* em cada aspecto linguístico) do LIWC em que os valores mais discrepantes encontram-se em negrito. Na referida tabela é possível observar que em alguns casos as traduções produzidas pelas ferramentas apresentaram perda de determinados aspectos linguísticos e por outras vezes apresentaram acréscimo dos mesmos em relação à tradução referência. Na busca por essas divergências em cada categoria, percebeu-se também que em diversos casos as três ferramentas produziram resultados muito diferentes umas das outras. Novamente, na maioria dos casos a ferramenta WL apresentou resultados mais distantes daqueles apresentados pelas ferramentas BI e GT.

Tabela VI.3: Médias de divergências de aspectos linguísticos verificados em TATs.

Categoria (Aspecto)	Tradução BI	Tradução GT	Tradução WL
`11-verb`	$+0.57\%$	-0.29%	**-5.16%**
`12-auxverb`	**+8.90%**	0.0%	$+5.95\%$
`14-present`	**+10.75%**	$+6.83\%$	$+3.05\%$
`13-past`	$+2.22\%$	$+2.22\%$	**-9.09%**
`15-future`	$+19.05\%$	**+29.17%**	$+19.05\%$
`16-adverb`	-2.22%	-4.44%	**+15.09%**
`10-article`	$+5.41\%$	$+8.50\%$	**+20.45%**
`17-preps`	-2.45%	-1.36%	**+4.68%**
`18-conj`	**+2.84%**	0.0%	$+0.49\%$
`5-we`	0.0%	0.0%	**+33.33%**
`6-you`	$+7.26\%$	$+7.26\%$	**+19.58%**
`7-shehe`	$+5.93\%$	$+4.31\%$	**-13.51%**
`8-they`	$+12.16\%$	**+15.58%**	$+9.72\%$
`2-pronoun`	$+6.35\%$	$+8.79\%$	**+2.78%**
`3-ppron`	$+2.87\%$	**+7.14%**	$+3.43\%$
`9-ipron`	$+8.04\%$	$+8.50\%$	**+3.43%**

Das três ferramentas analisadas, a ferramenta WL foi a que apresentou divergências mais significativas. Em relação à categoria `11-verb` (verbos) houve uma média de perdas significativa da ferramenta WL (**-5.16%**) contra uma média de perdas pouco significativa (-0.29%) do GT e um acréscimo médio pouco significativo ($+0.57\%$) do BI. A tabela VI.4 apresenta, em maiores detalhes, a sentença 11 do texto 7 em que as palavras “utilizando”, “poderia”, “aparecerem” e

“apareceriam” se associam à categoria `11-verb`. A palavra “acessar” não foi encontrada no LIWC em PB de 2007 e, portanto, foi contabilizada para a categoria `500-nfound`.

Tabela VI.4: Perdas na categoria `verbos` com o WL (Texto 7, sentença 11).

Sentença	**Perda de 52.98% na categoria `11-verb` com o WL**	**fp**	**rep**
Refer.	Com os **nanotransmissores** no lugar **utilizando somente o** pensamento você poderia **acessar** a internet e ao **invés das imagens aparecerem na** sua tela **elas apareceriam** dentro **da** sua mente.	4	3.36%
Cand. WL	Com os **nanotransmitters** no lugar **pelo** pensamento **sozinho** você poderia **logon ao Internet** e **em vez dos retratos que vêm acima em** sua tela **jogariam dentro de** sua mente.	2	1.58%

O WL também apresentou uma discrepância muito grande na categoria `13-past` (verbos no passado) com uma média de **-9.09%** contra $+2.22\%$ tanto para o BI quanto para o GT. Isso pode significar uma dificuldade do WL para lidar com a questão da flexão verbal. A tabela VI.5 traz como exemplo a sentença 7 do texto 10 em que as palavras “chutou”, “perdeu”, “foi”, “levado” e “mandado” foram associadas à referida categoria.

Tabela VI.5: Perdas na categoria `verbos no passado` com o WL (Texto 10, sentença 7)

Sentença	**Perda de 67.86% na categoria `13-past` com o WL**	**fp**	**rep**
Refer.	**Ele chutou a bola** e **perdeu**. **Ele** foi **levado teve os olhos vendados** e **foi mandado para** um **campo de prisioneiros** por três semanas.	6	5.88%
Cand. WL	**Retrocedeu** a **esfera** e **faltou-a**. Foi **removido blindfolded** e **emitido a** um **acampamento da prisão** por três semanas.	1	1.89%

A grande diferença verificada tanto na frequência de palavras (*i.e.,* de 6 para 1) quanto nos percentuais de representatividade (*i.e.,* de 5.88% para 1.89%) entre a sentença referência e a sentença candidata produzida pelo WL merecem destaque. As palavras “teve” e “retrocedeu” não foram encontradas no LIWC e, portanto, foram contabilizadas na categoria `500-nfound`.

Na categoria `7-shehe` (3ª pessoa do singular), o WL apresenta em média **-13.51%** contra $+5.93\%$ do BI e $+4.31\%$ do GT. Houve casos em que a perda em relação a essa categoria em determinadas sentenças produzidas pelo WL ficou entre 60.0% e 70.0% . Na sentença 15 do texto 6, por exemplo, essa perda chegou a 67.86%.

Na categoria `5-we` (1ª pessoa do plural), o WL apresenta um acréscimo acima do normal (média de **+33.33%**) enquanto as outras duas ferramentas não apresentam qualquer alteração. No entanto, ao verificar com mais atenção, chegou-se à conclusão de que esse acréscimo ocorreu em apenas 5 sentenças consideradas pequenas e que em apenas duas delas houve a utilização de uma palavra a mais contabilizada para essa categoria. Nos outros três casos tratou-se apenas de uma representatividade maior da referida categoria em relação às demais palavras utilizadas nas sentenças referência e candidata.

As ferramentas GT e BI também apresentaram anomalias menos significativas em relação a algumas categorias que refletem aspectos linguísticos. Cabe destacar que na categoria `3-ppron` (pronomes pessoais), o GT apresentou uma média de **+7.14%** contra médias de $+2.87\%$ do BI e de $+3.43\%$ do WL. Já na categoria `12-auxverb` (verbos auxiliares), o BI apresentou uma média de **+8.90%** contra 0.0% do GT e $+5.95\%$ do WL.

VI.4 Considerações sobre as divergências em aspectos psicolinguísticos

Nesta seção foi possível constatar a importância das novas perspectivas oferecidas pelo algoritmo `calcDPC`. Ao analisar de forma mais específica cada aspecto linguístico ou psicológico e as transformações desses aspectos provocadas por ferramentas de TAT, consegue-se ter indícios bem relevantes sobre as dificuldades apresentadas por cada ferramenta. Foi possível verificar, por exemplo, que em relação à categoria de palavras relacionadas à raiva, a ferramenta GT se sai muito bem enquanto as outras duas apresentam perdas significativas na média geral. Também foi possível fazer essa constatação por meio de uma análise mais pontual e detalhada, utilizando um exemplo real de uma sentença extraída de um dos dez textos.

As ferramentas BI e WL também apresentaram perdas consideráveis em relação à categoria que reflete aspectos familiares. De modo geral, a ferramenta WL mostrou um desempenho muito ruim em praticamente todas as categorias que refletem aspectos psicológicos e especialmente em categorias que refletem aspectos linguísticos. Categorias essas em que as ferramentas GT e BI se saem razoavelmente bem de modo geral. O GT, por exemplo, saiu-se muito bem em relação à categoria de verbos auxiliares.O desempenho do WL nas categorias de verbos e verbos no passado foi muito ruim, indicando que essa ferramenta apresenta dificuldades para realizar a flexão verbal. O desempenho do WL também foi muito ruim em relação à categoria que aloca palavras relacionadas à terceira pessoa do singular.

As informações apresentadas nesta seção podem ser de suma importância no processo de tomada de decisão sobre qual ferramenta utilizar para traduzir determinado tipo de texto. Se o texto expressa raiva, contém muitos aspectos familiares ou muitos verbos auxiliares, talvez a ferramenta GT seja a mais indicada. Se há necessidade de rigor com aspectos linguísticos há que se evitar o uso da ferramenta WL. Essas informações, além de auxiliarem o usuário final, também podem ajudar a desenvolver novas métricas e ajudar a melhorar a qualidade de determinadas ferramentas. Afinal, a partir desse algoritmo é possível observar como as ferramentas de TAT estão se comportando em relação à semântica (*e.g.,* aspectos linguísticos e psicológicos) no processo de tradução, abrindo uma série de possibilidades a serem exploradas. Outros exemplos específicos de perdas de determinados aspectos linguísticos ou psicológicos podem ser encontrado, em detalhes, na seção VIII.3 do anexo VIII deste trabalho.

Apesar da importante contribuição detalhada nesta seção, ao longo do presente trabalho ficou claro que o maior problema a ser resolvido, no entanto, é a ausência de métricas que sejam capazes de avaliar Traduções Automáticas de Texto (TATs) de forma satisfatória, com a observação da semântica presente nas sentenças avaliadas. As métricas conhecidas até então, incluindo o estado da arte (a métrica BLEU), se mostraram ineficientes na tarefa de considerar tais aspectos e tornar esse tipo de avaliação mais próxima da avaliação humana. Desta forma, fica claro que, em se tratando de métricas utilizadas para avaliar traduções, há uma demanda ainda não atendida. O desafio, a partir dos próximos experimentos, é verificar a eficácia da nova métrica no sentido de suprir essa demanda.

VI.5 Métrica BRAPT

Nesta seção são apresentados os experimentos realizados utilizando a métrica proposta (BRAPT) em comparação com o estado da arte (BLEU). Para estabelecer as comparações entre as métricas, foram selecionados dez textos jornalísticos escritos originalmente em inglês e traduzidos para o PB por dois profissionais especializados, que desse ponto em diante passam a ser referenciados como `especialista 1` e `especialista 2`, produzindo o que se conhece por traduções referência. Em seguida, os mesmos textos foram submetidos a três ferramentas de TAT conhecidas em ambiente *web*: Google Tradutor (GT)[1], Bing Tradutor (BI)[2] e WorldLingo Tradutor (WL)[3], produzindo o que se conhece por traduções candidatas.

As avaliações experimentais foram organizadas em cinco partes. Na primeira parte, indicada na subseção VI.5.1, é realizada uma análise comparativa das traduções produzidas pelas ferramentas GT, BI e WL em comparação com as traduções produzidas por um especialista. Nessas análises foram aplicadas as métricas BLEU e BRAPT a fim de verificar se ambas apresentam resultados consistentes sobre o desempenho das ferramentas. Essa fase dos experimentos também vem acompanhada das impressões de um dos especialistas, que a partir deste ponto passa a ser referenciado como `especialista avaliador`. Na subseção VI.5.2 as métricas são aplicadas novamente às mesmas traduções comparadas com traduções realizadas pelo `especialista 1` e também pelo `especialista 2` objetivando verificar se há diferenças significativas ao realizar comparações com traduções realizadas por humanos distintos. Em seguida, a subseção VI.5.3, apresenta uma análise comparativa dos índices de compatibilidade medidos pelas métricas BLEU e BRAPT frente às avaliações do `especialista avaliador`. A subseção VI.5.4, apresenta uma análise detalhada sobre algumas sentenças, facilitando a compreensão da compatibilidade verificada pelas métricas BLEU e BRAPT. Finalmente, na subseção VI.5.5 são relatadas algumas limitações da métrica BRAPT identificadas durante os experimentos.

VI.5.1 Análise comparativa das ferramentas de TAT

Nesta subseção foi realizada uma análise texto a texto utilizando a métrica BLEU e a métrica proposta no presente trabalho (BRAPT). Essas métricas foram utilizadas para comparar as traduções referência produzidas pelo `especialista 1` com as traduções candidatas produzidas pelas ferramentas BI, GT e WL. O objetivo consiste em avaliar a consistência das duas métricas ao serem aplicadas em resultados provenientes dessas ferramentas cujo desempenho também é avaliado. Os gráficos apresentados nas figuras VI.3 e VI.4 devem ajudar nessa análise.

[1] https://www.google.com.br/translator
[2] https://www.bing.com/translator/
[3] http://http://www.worldlingo.com/

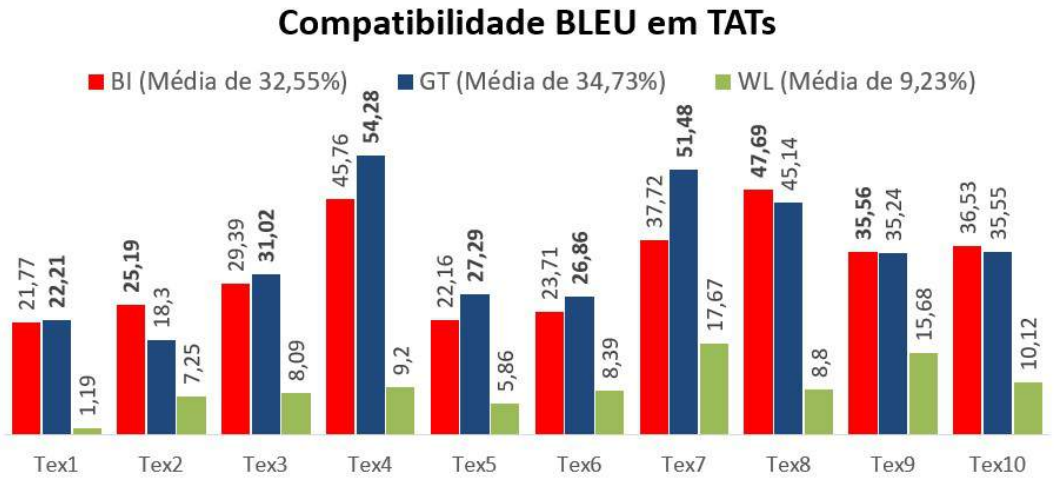


Figura VI.3: Compatibilidade média dos textos de acordo com a BLEU.

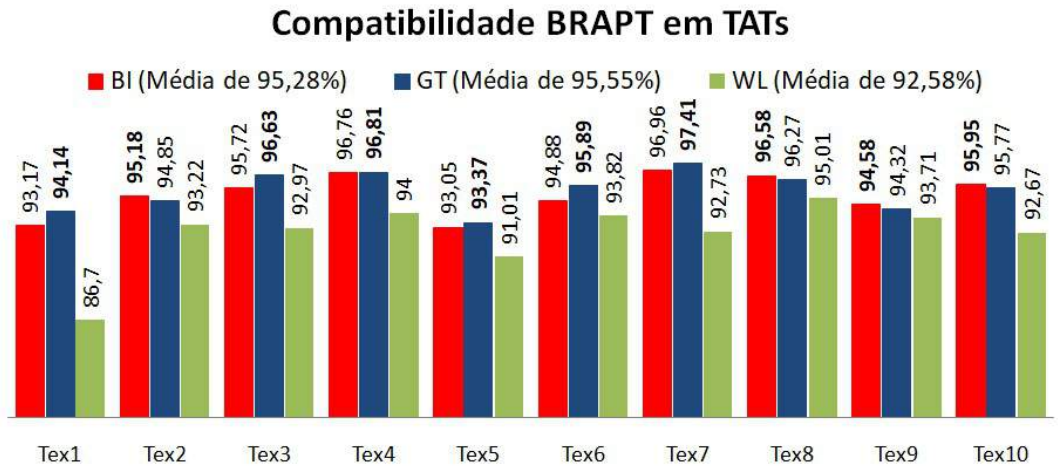


Figura VI.4: Compatibilidade média dos textos de acordo com a BRAPT.

Os gráficos apresentam resultados consistentes, visto que em ambas as métricas é possível observar que as ferramentas GT e BI apresentam um desempenho parelho, com traduções candidatas mais semelhantes às traduções referência, e bem superior ao desempenho da ferramenta WL em cada texto e na média geral. Até mesmo na ordem de desempenho de cada ferramenta os resultados são iguais, incluindo a média geral de cada ferramenta. As duas métricas apresentam um mesmo diagnóstico sobre o desempenho das ferramentas em todos textos.

Em 114 das 128 sentenças presentes nos dez textos, a métrica BRAPT indica que as ferramentas GT e BI apresentaram traduções candidatas mais similares às traduções referências do que a ferramenta WL. Segundo a BRAPT, o GT apresenta melhor desempenho em 42% das traduções enquanto o BI apresenta melhor desempenho em 41% das traduções. Houve empate entre o GT e o BI em 6% das traduções e o WL obteve melhor desempenho em apenas 11% das traduções. Os dados apresentados são consistentes com o que existe na literatura sobre a análise das ferramentas BI e GT, haja visto os resultados apresentados em [de Melo et al., 2015]. Esses dados podem ser observados na figura VI.5, que sinaliza vantagem discreta para o GT.

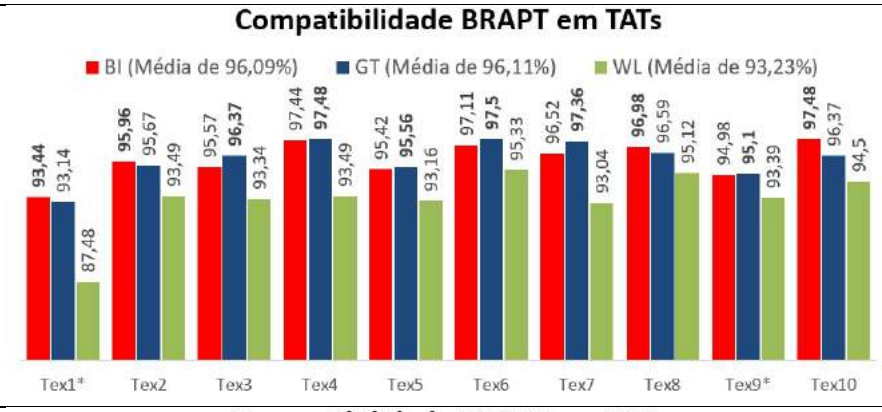
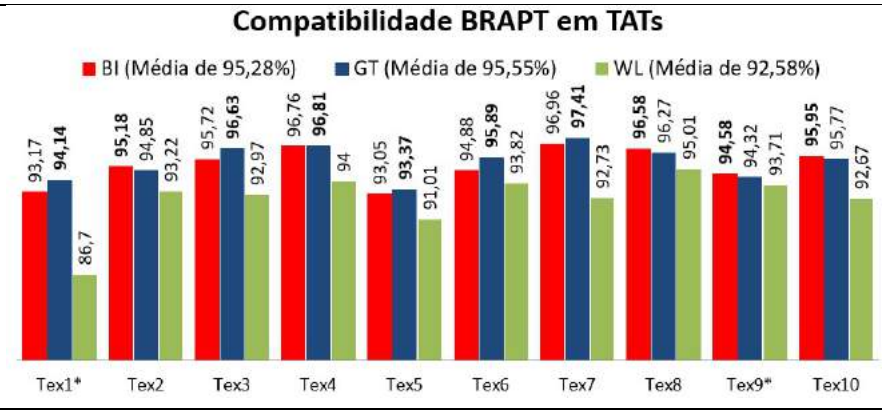
Figura VI.5: Desempenho das ferramentas de TAT em sentenças, de acordo com a BRAPT.

Os textos produzidos pelas ferramentas de TAT também foram analisados pelo `especialista avaliador` que considerou as traduções do WL bem inferiores às traduções produzidas pelo BI e pelo GT. Esses resultados também mostram-se consistentes, visto que coincidem com os resultados da aplicação da métrica BRAPT.

VI.5.2 Consistência entre as traduções dos especialistas

As métricas BLEU e BRAPT foram aplicadas novamente tendo como referência, desta vez, as traduções produzidas pelo `especialista 2`. Esses resultados foram comparados com os resultados alcançados com as traduções do `especialista 1` e apresentados na tabela VI.6.

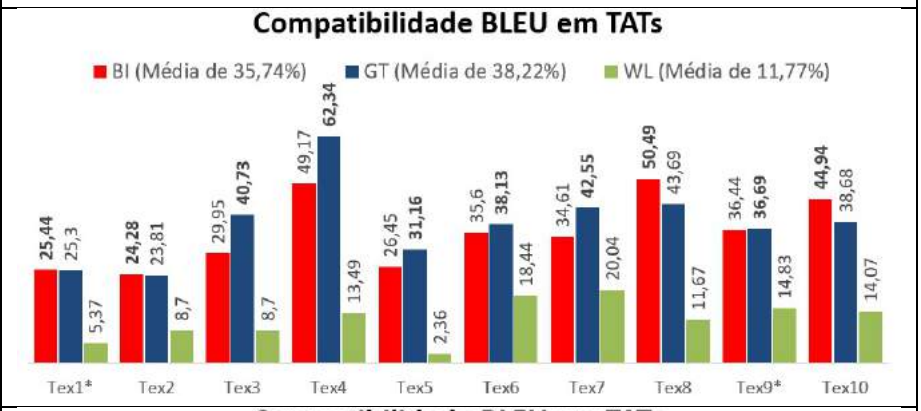
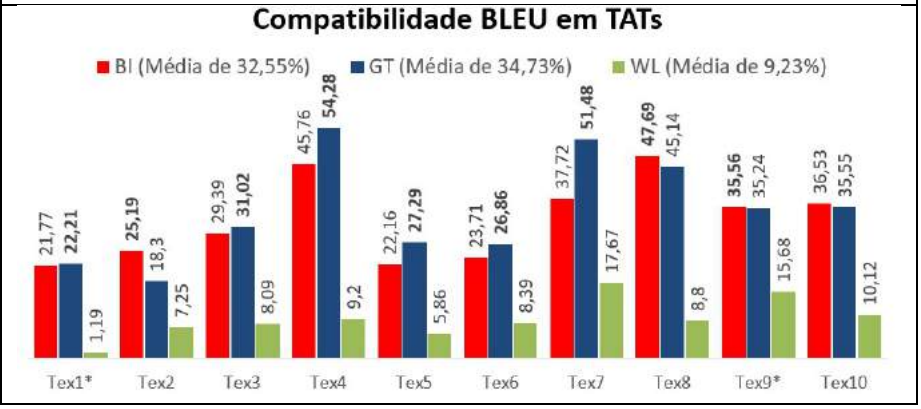
Tabela VI.6: Métrica BRAPT: Comparação entre especialistas.

Tradutor	Gráfico
Especialista 2	
Especialista 1	

O objetivo é verificar se há diferenças significativas ao aplicar as referidas métricas a traduções realizadas por especialistas distintos. Como pode ser visto, em oito textos, os resultados em relação à ordem de desempenho das três ferramentas são exatamente os mesmos. As diferenças, que são sutis, ficam por conta dos percentuais. Apenas em dois dos dez textos (*i.e.,* `Tex1*` e `Tex9*`), houve uma simples inversão na ordem de desempenho entre as ferramentas BI e GT. Nos dez textos essas duas ferramentas apresentaram desempenho parelho e superior ao da ferramenta WL tanto na comparação com as traduções produzidas pelo `especialista 1`, quanto na comparação com as traduções produzidas pelo `especialista 2`. Em ambos os experimentos, a ferramenta GT obteve melhor desempenho em seis textos e a ferramenta BI em quatro textos.

Ao realizar as mesmas comparações com a aplicação da métrica BLEU, novamente houve apenas pequenas diferenças quanto aos percentuais. Os resultados também foram exatamente iguais aos produzidos com a aplicação da métrica BRAPT considerando a ordem de desempenho das três ferramentas. Mais uma vez, assim como ocorreu em relação à BRAPT, houve apenas uma inversão na ordem de desempenho das ferramentas BI e GT nos mesmos textos `Tex1*` e `Tex9*`. Nos demais textos, bem como nas médias, não houve qualquer alteração em relação aos testes com a BRAPT, incluindo o melhor desempenho da ferramenta GT nos mesmos seis textos e da ferramenta BI nos mesmos quatro textos. Os resultados podem ser vistos na tabela VI.7.

Tabela VI.7: Métrica BLEU: Comparação entre especialistas.

Tradutor	**Gráfico**
Especialista 2	
Especialista 1	

Como já era esperado, humanos, ao realizarem uma tradução, dificilmente produzem resultados exatamente iguais. Isso ocorre por que cada pessoa possui um idioleto (*i.e.,* uma maneira própria de utilizar recursos linguísticos) baseado em suas próprias experiências. Devido a essa maneira particular com que cada indivíduo se apropria de recursos linguísticos, no momento da tradução dois profissionais podem escolher utilizar palavras distintas (*e.g.,* "odiar" ou "detestar", "bela" ou "bonita") para representar um mesmo significado [Brandão, 2009, JOHASSON, 2004].

Contudo, como pôde ser verificado nas comparações, apesar de não serem exatamente iguais, os resultados são semelhantes e não apresentam diferenças significativas em relação à utilização de traduções referência produzidas por especialistas distintos. Por essa razão, a partir deste ponto o presente trabalho volta-se novamente apenas para as traduções referência produzidas por um único especialista a fim de evitar a exibição de resultados repetitivos.

VI.5.3 Análise do índice de compatibilidade

Na subseção anterior as métricas BLEU e BRAPT apresentaram resultados bem consistentes em relação ao desempenho das ferramentas de TAT. Apesar disso, houve uma disparidade perceptível em relação à valoração das compatibilidades apresentadas pelas duas métricas. Esses resultados motivam a seguinte pergunta: qual dessas métricas se aproxima mais de uma avaliação humana? Com o intuito de responder a essa pergunta, as compatibilidades apresentadas pelas duas métricas foram comparadas com as impressões de um `especialista avaliador`, que utilizou uma escala (*rating scale*) de 0 a 10 para avaliar cada texto traduzido pela ferramenta GT, que foi a que apresentou o melhor desempenho nos testes anteriores. Ao observar a figura VI.6 percebe-se, de forma inequívoca, que a métrica BRAPT (representada pelas barras azuis) apresenta percentuais de compatibilidade bem mais próximos das avaliações realizadas pelo `especialista avaliador` (representadas pelas barras pretas) em todos os dez textos.

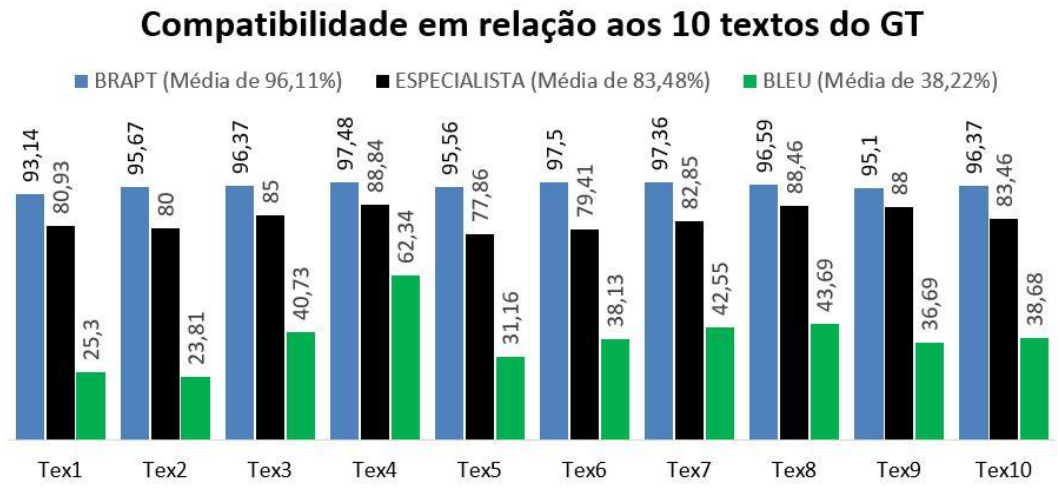


Figura VI.6: Compatibilidade verificada em textos traduzidos pelo GT.

Notadamente a métrica BLEU (representada pelas barras verdes) apresenta percentuais bem mais baixos em todos os textos. Com exceção do texto 4, os percentuais apresentados pela métrica BLEU correspondem quase sempre a menos da metade dos percentuais apresentados pela métrica BRAPT e pelas avaliações do `especialista avaliador`. Nos textos 1 e 3, por exemplo, essa diferença chega a representar menos de um terço na comparação com os valores produzidos pela métrica BRAPT e pelas avaliações do `especialista avaliador`.

Objetivando uma confirmação dos resultados e até mesmo uma análise mais detalhada, a mesma comparação foi realizada com as sentenças do texto 6 traduzidas pelo GT. Esse texto foi escolhido por se tratar do texto com o maior número de sentenças (17). Os resultados são apresentados pela figura VI.7.

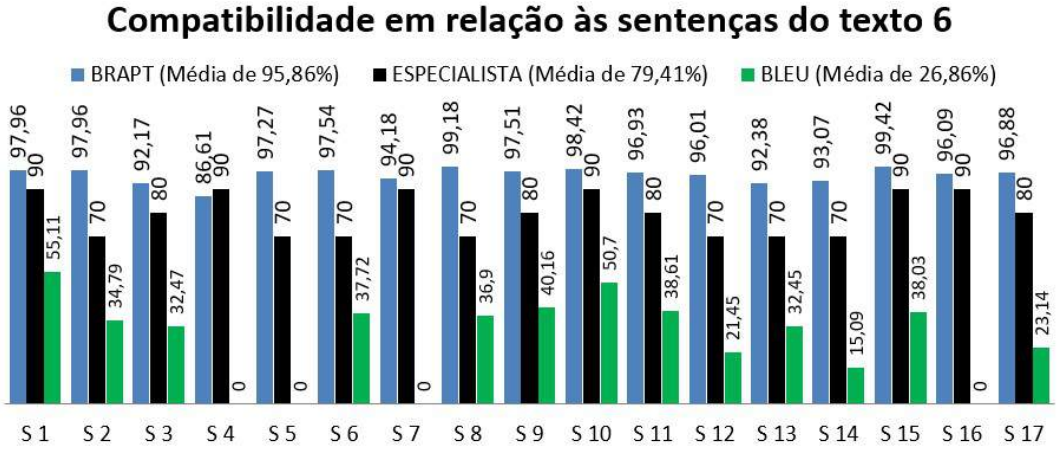


Figura VI.7: Compatibilidade verificada em sentenças traduzidas pelo GT.

Novamente a métrica BLEU apresentou percentuais bem inferiores aos percentuais apresentados pela métrica BRAPT e pelas avaliações do `especialista avaliador`. Vale destacar a incompatibilidade total de algumas sentenças segundo a métrica BLEU (*i.e.,* `S4`, `S5`, `S7` e `S16`). Dentre essas quatro sentenças citadas, a menor (`S7`) e a maior sentença (`S16`) são apresentadas nessa ordem para uma análise minuciosa. Os resultados são exibidos na figura VI.8.

Referência: Mas não **o** são. **Candidata (GT):** Mas **eles** não são.		
Especialista: 90%	**BRAPT:** 94,2%	**BLEU:** 0%

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.		
Especialista: 90%	**BRAPT:** 96,1%	**BLEU:** 0%

Figura VI.8: Sentenças 7 e 16 do texto 6 produzidas pelo GT: BLEU vs BRAPT.

A figura acima traz as sentenças referência e candidata, além das avaliações do `especialista`

`avaliador` e das compatibilidades BRAPT e BLEU. Cabe, no entanto, ratificar que para a métrica BRAPT a ordem das palavras não possui relevância. As sentenças em questão explicitam o rigor da métrica BLEU ao descartar (*i.e.,* compatibilidade = 0%) sentenças candidatas com significados próximos aos da sentença referência. Na métrica BRAPT, bem como na visão do `especialista avaliador`, essas variações caracterizam apenas uma redução no percentual de compatibilidade ao invés de representar uma incompatibilidade total das referidas sentenças.

VI.5.4 Análise de situações específicas em sentenças traduzidas

Também foram realizados experimentos com sentenças avaliadas de forma ineficaz pela métrica BLEU. Esses experimentos visaram analisar os seguintes aspectos: (a) a substituição de uma palavra por um sinônimo; (b) a inversão de ordem das palavras e (c) a avaliação de candidatas menores com substituição de dois termos por um sinônimo (*e.g.,* "muito grande" por "enorme"). Novamente a métrica BRAPT apresentou percentuais mais condizentes com o significado das sentenças referência e candidata avaliadas. Aa figura VI.9, que contém sentenças que representam os aspectos a,b e c (*i.e.,* sentenças 1, sentenças 2 e sentenças 3), em maiores detalhes.

Referência 1: Ela vai **odiar** meu novo carro		
Candidata 1: Ela vai **detestar** meu novo carro		
Especialista: 100%	**BRAPT:** 93,8%	**BLEU:** 0%
Referência 2: Minha **nova namorada** é muito alta		
Candidata 2: Minha **namorada nova** é muito alta		
Especialista: 100%	**BRAPT:** 100%	**BLEU:** 0%
Referência 3: A casa é **muito grande**		
Candidata 3: A casa é **enorme**		
Especialista: 90%	**BRAPT:** 79.21%	**BLEU:** 0%

Figura VI.9: Compatibilidade BLEU VS BRAPT em situações específicas.

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 referida métrica apresentou incompatibilidade total (0%). Em geral, essa compatibilidade 0 da métrica BLEU se dá em decorrência de incompatibilidade total em algum nível de *n-gramas*, de forma mais recorrente nos *4-gramas* (*i.e.,* tetragramas).

Nas três subseções a seguir, são apresentados ilustrações e cálculos referentes à aplicação da métrica BLEU a fim de identificar as causas dessa ausência de compatibilidade nas três sentenças. Também serão exibidos dados que ajudem a entender a razão pela qual a métrica BRAPT apresenta um desempenho melhor nesses casos.

Substituição de uma palavra por um sinônimo

Uma peculiaridade da métrica BLEU, visível na figura VI.10, é que se houver o comprometimento total de um único grau de *n-grama*, como foi o caso dos tetragramas, automaticamente a sentença fica completamente comprometida, resultando em 0% de compatibilidade. Nessa ilustração e em todos os próximos exemplos de funcionamento da métrica BLEU, as incompatibilidades estão destacadas em amarelo e resultam em diminuição na precisão modificada de *n-gramas* (coluna "p-n").

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	

Figura VI.10: Substituição de uma palavra por um sinônimo de acordo com a BLEU.

A figura VI.11 traz um exemplo da aplicação do cálculo da compatibilidade BLEU em detalhes.

$$BLEU(S_{Ref1}, S_{Cand1}) = BP(S_{Ref1}, S_{Cand1}) \cdot (\frac{5}{6} \cdot \frac{3}{5} \cdot \frac{1}{4} \cdot \frac{0}{3})^{\frac{1}{4}}$$

$$BLEU(S_{Ref1}, S_{Cand1}) = 1 \cdot (0{,}83 \cdot 0{,}6 \cdot 0{,}25 \cdot 0)^{0{,}25} = 0{,}0$$

Figura VI.11: Aplicação da métrica BLEU na substituição de uma palavra por um sinônimo.

Nesse exemplo, a compatibilidade BRAPT computada foi de 93.8%. As categorias identificadas nas duas únicas palavras divergentes entre as sentenças referência e candidata (*i.e.,* "odiar" e "detestar") são exibidas na tabela VI.8. A única categoria divergente (*i.e.,* `124-humans`) encontra-se sublinhada. A título de informação, além da palavra "detestar", a palavra "novo", contida em ambas as sentenças, também se enquadra na categoria `124-humans`, assim como ocorre na versão do LIWC de 2007 para a língua inglesa.

Tabela VI.8: Métrica BRAPT: Categorias relacionadas às palavras "detestar" e "odiar".

Palavras	**Categorias BRAPT**
detestar	`11-verb, 125-affect, 127-negemo, 129-anger, 124-humans`
odiar	`11-verb, 125-affect, 127-negemo, 129-anger`

As figuras VI.12 e VI.13 representam, respectivamente, os vetores referência ($\vec{v}_{Ref1}$) e candidato ($\vec{v}_{Cand1}$) estudados nesse exemplo. Os referidos vetores utilizados para o cômputo da

compatibilidade BRAPT apresentam somente algumas das principais categorias.

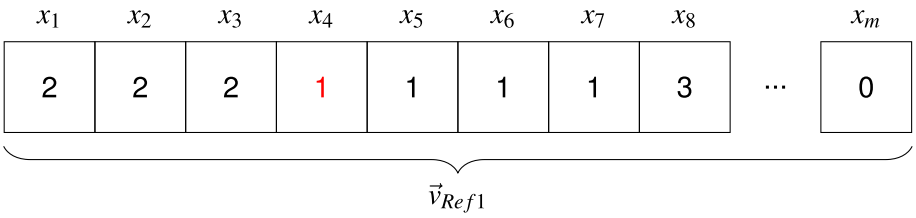


Figura VI.12: Representação vetorial da sentença referência 1.

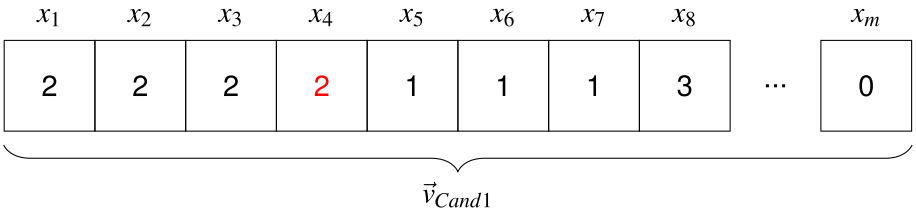


Figura VI.13: Representação vetorial da sentença candidata 1.

A tabela VI.9 exibe a categoria representada por cada posição dos referidos vetores. A categoria em que houve diferenças na frequência de palavras está destacada em vermelho.

Tabela VI.9: Detalhamento dos vetores $\vec{v}_{Ref1}$ e $\vec{v}_{Cand1}$

Posição	**Categoria**	**Descrição**
x_1	`11-verb`	Verbos
x_2	`2-pronoun`	Pronomes
x_3	`3-ppron`	Pronomes pessoais
x_4	`124-humans`	Aspectos humanos
x_5	`125-affect`	Afeto
x_6	`127-negemo`	Emoções negativas
x_7	`129-anger`	Raiva
x_8	`250-relativ`	Relatividade
x_9 a x_{m-1}	`...`	Demais categorias
x_m	`500-nfound`	Palavras não encontradas

As figuras VI.14 e VI.15 apresentam as duas sentenças por meio de dois vetores de representatividade percentual (*i.e.,* $\vec{v}_{RefP1}$ e $\vec{v}_{CandP1}$), gerados conforme detalhado no capítulo IV. Nesses novos vetores é possível observar, por exemplo, que na sentença referência 1, que teve 28 palavras contabilizadas no total, a categoria `124-humans` é menos representativa (3.57%) do que na sentença candidata 1, que teve 29 palavras contabilizadas. Na candidata 1 a categoria `124-humans` teve uma representatividade de 6.9% ($\frac{2}{29} \cdot 100 = 6.9$). Também é possível notar que, dentre as categorias ilustradas, a categoria com o maior percentual de representatividade em ambos os vetores é a categoria `250-relativ` (*i.e.,* a categoria que reflete aspectos de relatividade).

Por fim, observando os dois novos vetores e a tabela VI.10, é possível identificar o percentual de representatividade de cada categoria tanto na sentença referência ($\vec{v}_{RefP1}$) quanto na

sentença candidata ($\vec{v}_{CandP1}$). Também é importante observar que, exceto pela posição x_m, houve divergências em relação à representatividade de todas as categorias ilustradas.

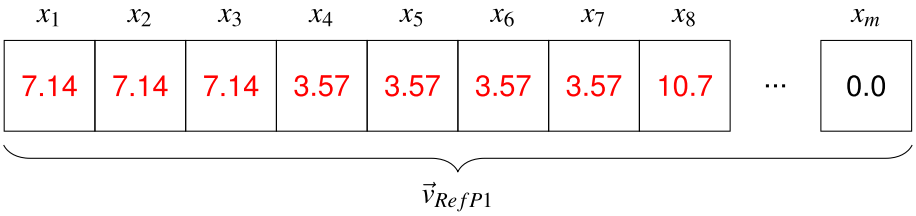


Figura VI.14: Representatividade percentual da sentença referência 1.

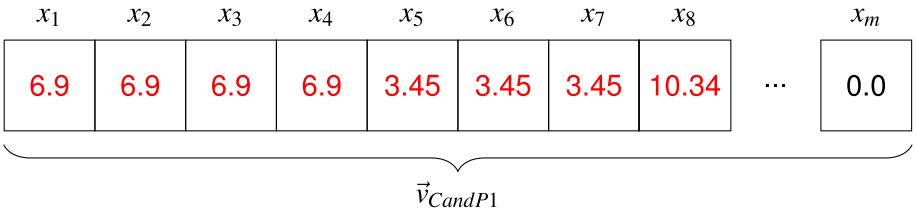


Figura VI.15: Representatividade percentual da sentença candidata 1.

Tabela VI.10: Detalhamento dos vetores $\vec{v}_{RefP1}$ e $\vec{v}_{CandP1}$

Posição	**Categoria**	**Descrição**	**Repres.** $\vec{v}_{RefP1}$	**Repres.** $\vec{v}_{CandP1}$
x_1	11-verb	Verbos	7.14 %	6.9 %
x_2	2-pronoun	Pronomes	7.14 %	6.9 %
x_3	3-ppron	Pronomes pessoais	7.14 %	6.9 %
x_4	124-humans	Aspectos humanos	3.57 %	6.9 %
x_5	125-affect	Afeto	3.57 %	3.45 %
x_6	127-negemo	Emoções negativas	3.57 %	3.45 %
x_7	129-anger	Raiva	3.57 %	3.45 %
x_8	250-relativ	Relatividade	10.7 %	10.34 %
x_9 a x_{m-1}	...	Demais categorias	...	...
x_m	500-nfound	Palavras não encontradas	0.0 %	0.0 %

Inversão de ordem de palavras

A figura VI.16, ilustra outra sentença em que os tetragramas são afetados e, consequentemente, comprometem a compatibilidade BLEU. Como nesse exemplo duas palavras ficaram fora do padrão, o impacto na precisão de *n-gramas* foi ainda maior e gradativo a partir dos bigramas. A figura VI.17 apresenta o cálculo da compatibilidade BLEU para essa simples inversão de duas palavras. As palavras envolvidas nas duas sentenças são exatamente as mesmas e não há comprometimento significativo da sentença candidata em relação à sentença referência.

A métrica BRAPT não considera a ordem das palavras. Logo, os dois vetores produzidos são exatamente iguais. Portanto, no exemplo acima e, consequentemente, na aplicação dos referidos cálculos, o cômputo da compatibilidade BRAPT é de 100% contra 0% da compatibilidade BLEU.

n-gram	sent.	Compatibilidade BLEU (Referência vs Candidata)										p-n
1-gram	Ref.	Minha	nova	namorada	é	muito	alta					6/6
	Cand.	Minha	namorada	nova	é	muito	alta					
2-gram	Ref.	Minha nova		nova namorada		namorada é		é muito		muito alta		2/5
	Cand.	Minha namorada		namorada nova		nova é		é muito		muito alta		
3-gram	Ref.	Minha nova namorada			nova namorada é			namorada é muito			é muito alta	1/4
	Cand.	Minha namorada nova			namorada nova é			nova é muito			é muito alta	
4-gram	Ref.	Minha nova namorada é				nova namorada é muito				namorada é muito alta		0/3
	Cand.	Minha namorada nova é				namorada nova é muito				nova é muito alta		

Figura VI.16: Inversão de ordem das palavras de acordo com a métrica BLEU.

$$BLEU(S_{Ref2}, S_{Cand2}) = BP(S_{Ref2}, S_{Cand2}) \cdot (\frac{6}{6} \cdot \frac{2}{5} \cdot \frac{1}{4} \cdot \frac{0}{3})^{\frac{1}{4}}$$

$$BLEU(S_{Ref2}, S_{Cand2}) = 1 \cdot (1 \cdot 0{,}6 \cdot 0{,}4 \cdot 0)^{0{,}25} = 0{,}0$$

Figura VI.17: Exemplo de aplicação da métrica BLEU na inversão de ordem de palavras.

Substituição de duas palavras por um sinônimo

A figura VI.18 ilustra a aplicação da métrica BLEU na substituição de duas palavras por um sinônimo, além da aplicação BP. Essa penalidade é aplicada sempre que a sentença candidata é menor do que a referência. Além das incompatibilidades já comentadas nos exemplos anteriores, é interessante observar, em todos os níveis, a ausência de um último *n-grama* na sentença candidata para ser comparado com um *n-grama* presente na sentença referência. Isso gera comprometimento da precisão de *n-gramas* em todos os níveis, além da multiplicação pela BP, que nesse caso é o exponencial de $1 - \frac{r}{c}$ (*i.e.,* o exponencial de $1 - \frac{5}{4}$).

n-gram	sent.	Compatibilidade BLEU (Referência vs Candidata)									p-n
1-gram	Ref.	A	casa	é	muito	grande					3/5
	Cand.	A	casa	é	enorme						
2-gram	Ref.	A casa		casa é		é muito		muito grande			2/4
	Cand.	A casa		casa é		é enorme					
3-gram	Ref.	A casa é			casa é muito			é muito grande			1/3
	Cand.	A casa é			casa é enorme						
4-gram	Ref.	A casa é muito				casa é muito grande					0/2
	Cand.	A casa é enorme									

Figura VI.18: Candidata menor de acordo com a métrica BLEU.

Apesar da penalidade por brevidade (BP) gerar uma punição extra, mais uma vez, o comprometimento total de um nível de *n-grama* (*i.e.,* dos tetragramas) foi o fator determinante para que a compatibilidade resultasse em 0%. A figura VI.19 traz o cálculo da compatibilidade BLEU para esse exemplo. Em relação à métrica BRAPT. a compatibilidade não foi alta (79.21%), mas ao menos não representou um descarte completo da sentença candidata.

As figuras VI.20 e VI.21 representam, respectivamente, os vetores referência $(\vec{v}_{Ref3})$ e candidato $(\vec{v}_{Cand3})$ estudados nesse exemplo. Novamente, esses vetores apresentam somente algumas

das principais categorias presentes em ambas as sentenças.

$$BLEU(S_{Ref3}, S_{Cand3}) = BP(S_{Ref3}, S_{Cand3}) \cdot (\frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{0}{2})^{\frac{1}{4}}$$

$$BLEU(S_{Ref3}, S_{Cand3}) = e^{(1-\frac{5}{4})} \cdot (0,6 \cdot 0,5 \cdot 0,33 \cdot 0)^{0,25} = 0,0$$

$$BLEU(S_{Ref3}, S_{Cand3}) = 0,78 \cdot (0,6 \cdot 0,5 \cdot 0,33 \cdot 0)^{0,25} = 0,0$$

Figura VI.19: Exemplo de aplicação da BLEU na substituição de duas palavras por um sinônimo.

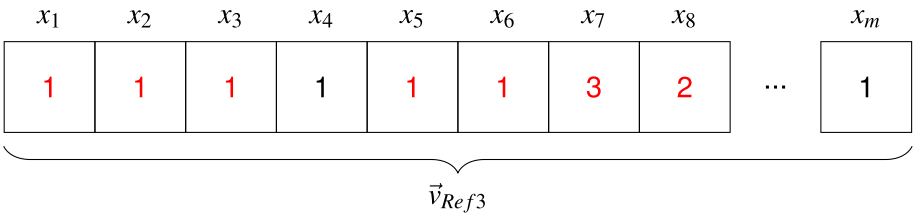


Figura VI.20: Representação vetorial da sentença referência 3.

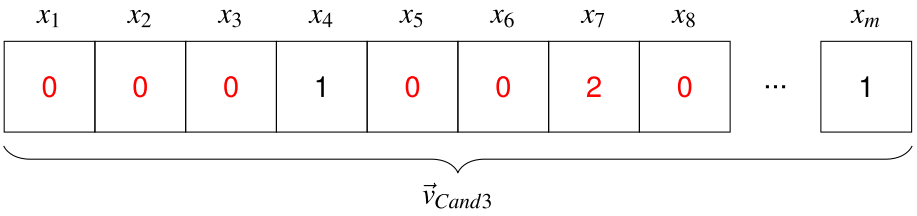


Figura VI.21: Representação vetorial da sentença candidata 3.

Observando os dois vetores ($\vec{v}_{Ref3}$ e $\vec{v}_{Cand3}$) é possível perceber que houve frequências diferentes (texto destacado em vermelho) de palavras em sete categorias, algumas delas já esperadas como as categorias `16-adverb` (posição x_1), `20-quant` (posição x_2) e `250-relativ` (posição x_7). Outras categorias, como `252-space` (posição x_4) e `500-nfound` (posição x_m) não apresentaram diferenças na frequência de palavras. No entanto, cabe lembrar que, conforme apresentado na subseção IV.1.2, embora a frequência seja a mesma no caso de algumas posições (*e.g.,* x_m), os pesos são diferentes, visto que $\vec{v}_{Ref3}$ possui 31 contabilizações e $\vec{v}_{Cand3}$ possui apenas 14 contabilizações. Logo, em sua versão normalizada, a categoria representada por x_m possui, na verdade, valores distintos. Esses valores são 3.22 (*i.e.,* $\frac{1}{31}$) e 7.14 (*i.e.,* $\frac{1}{14}$) para as sentenças referência e candidata, respectivamente. O mesmo se aplica às demais categorias. Não se pode confundir frequência de palavras em determinada categoria com a representatividade da categoria para a referida sentença.

As sentenças envolveram seis palavras distintas (*i.e.,* “A”, “casa”, “é”, “muito”, “grande” e “enorme”) que foram contabilizadas em 23 das 65 categorias contidas na BRAPT. A sentença referência 3 teve 31 palavras contabilizadas em 23 categorias distintas, enquanto a sentença candidata 3 teve apenas 14 contabilizações em 12 categorias distintas. A tabela VI.11 traz a categoria

representada por cada posição dos referidos vetores (*i.e.,* $\vec{v}_{Ref3}$ e $\vec{v}_{Cand3}$) e as figuras VI.22 e VI.23 apresentam as duas sentenças por meio de dois vetores de representatividade percentual (*i.e.,* $\vec{v}_{RefP3}$ e $\vec{v}_{CandP3}$).

Tabela VI.11: Frequência de palavras nos vetores $\vec{v}_{Ref3}$ e $\vec{v}_{cand3}$

Posição	**Categoria**	**Descrição**	$\vec{v}_{Ref3}$	$\vec{v}_{cand3}$
x_1	`16-adverb`	Advérbios	1	0
x_2	`20-quant`	Quantidade	1	0
x_3	`134-discrep`	Discrepância	1	0
x_4	`252-space`	Aspectos relativos a espaço	1	1
x_5	`136-certain`	Certeza	1	0
x_6	`131-cogmech`	Aspectos cognitivos	1	0
x_7	`250-relativ`	Relatividade	3	2
x_8	`150-ingest`	Ingestão	2	0
x_9 a x_{m-1}	`...`	Demais categorias	19	10
x_m	`500-nfound`	Palavras não encontradas	1	1

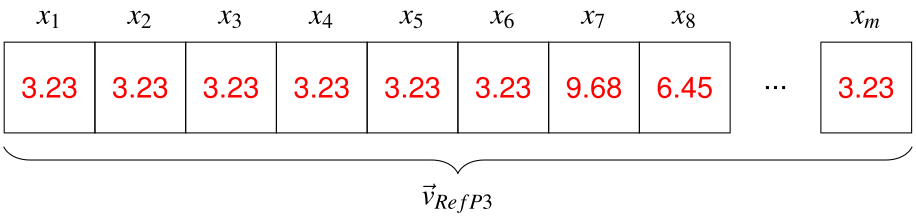


Figura VI.22: Representatividade percentual da sentença referência 3.

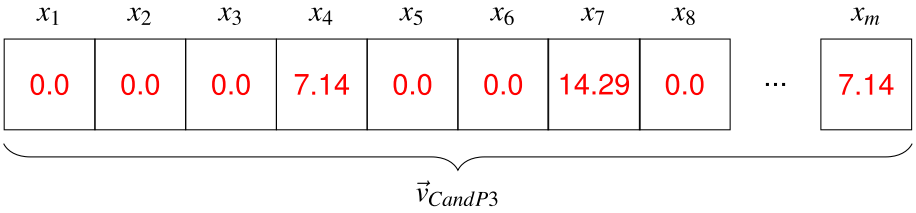


Figura VI.23: Representatividade percentual da sentença candidata 3.

Nesses novos vetores é importante frisar que agora as divergências (destacadas em vermelho) ficaram evidenciadas em todas as posições. Também é relevante observar, por exemplo, que na sentença candidata 3, diversas categorias não foram contabilizadas: `16-adverb`, `20-quant`, `134-discrep`, `136-certain`, `131-cogmech` e `150-ingest`. A categoria mais representativa em ambas as sentenças é a categoria `250-relativ`, que teve respectivamente 9.68% e 14.29% de representatividade em $\vec{v}_{RefP3}$ e $\vec{v}_{CandP3}$. Tanto na sentença referência 3 quanto na sentença candidata 3, a categoria `252-space` (posição x_4) teve apenas uma palavra contabilizada. Contudo, em $\vec{v}_{CandP3}$ essa categoria é muito mais representativa (com 7.14% contra 3.23% em $\vec{v}_{RefP3}$). Isso ocorre por que em $\vec{v}_{CandP3}$ essa categoria representa uma de 14 partes (*i.e.,* contabilizações). Já em $\vec{v}_{RefP3}$, ela representa uma de 31 partes. Outras quatorze categorias que não aparecem na

ilustração também apresentaram representatividades diferentes entre as sentenças referência e candidata. A tabela VI.12 apresenta o percentual de representatividade de cada categoria tanto na sentença referência ($\vec{v}_{RefP3}$) quanto na sentença candidata ($\vec{v}_{CandP3}$). As divergências novamente estão destacadas em vermelho.

Tabela VI.12: Representatividade das categorias dos vetores $\vec{v}_{RefP3}$ e $\vec{v}_{CandP3}$.

Posição	Categoria	Descrição	$\vec{v}_{RefP3}$	$\vec{v}_{CandP3}$
x_1	`16-adverb`	Advérbios	3.23 %	0.0 %
x_2	`20-quant`	Quantidades	3.23 %	0.0 %
x_3	`134-discrep`	Discrepância	3.23 %	0.0 %
x_4	`252-space`	Aspectos relativos a espaço	3.23 %	7.14 %
x_5	`136-certain`	Certeza	3.23 %	0.0 %
x_6	`131-cogmech`	Aspectos cognitivos	3.23 %	0.0 %
x_7	`250-relativ`	Relatividade	9.68 %	14.29 %
x_8	`i50-ingest`	Ingestão	6.45 %	0.0 %
x_9 a x_{m-1}	`...`	Demais categorias	...	...
x_m	`500-nfound`	Palavras não encontradas	3.23 %	7.14 %

VI.5.5 Limitações da métrica BRAPT

Durante os experimentos foram utilizadas as métricas BRAPT e BLEU com o intuito de comparar seus desempenhos, além de avaliar as três ferramentas de TAT estudadas. Objetivando dar mais consistência aos experimentos, também foi solicitado a um especialista (*i.e.,* o `especialista avaliador`) que fizesse suas considerações acerca de todas as 384 sentenças traduzidas (*i.e.,* 128 sentenças produzidas por cada uma das três ferramentas estudadas).

Em relação às ferramentas, tanto as compatibilidades BLEU e BRAPT quanto as considerações do `especialista avaliador` indicaram que o Google Tradutor (GT) e o Bing Tradutor (BI) se mostraram mais sofisticadas, com resultados mais próximos da avaliação humana, do que a ferramenta World Lingo (WL). Na comparação entre as métricas, os resultados indicaram que a BRAPT apresenta percentuais de compatibilidade bem mais próximos da avaliação humana do que a BLEU praticamente em todas as sentenças produzidas pelas ferramentas BI e GT. O mesmo ocorreu em relação à ferramenta World Lingo (WL), ainda que em menor escala.

Sobre ferramenta WL, de acordo com as análises do `especialista avaliador`, houve casos em que a BLEU se saiu melhor do que a BRAPT, especialmente ao apontar 0% de compatibilidade para traduções que, de fato, não conseguem transmitir minimamente o conteúdo original contido na sentença referência. A figura VI.24 traz alguns exemplos de sentenças retiradas dos dez textos estudados. Cada sentença aparece acompanhada das compatibilidades BRAPT e BLEU, além da avaliação do `especialista avaliador`.

Referência 4: Até recentemente tais **buscas estavam no domínio de vigaristas e charlatões**. **Candidata 4:** Até recentemente tais **perseguições eram o reino dos quacks e charlatans**.		
Especialista: 0%	**BRAPT:** 83.9%	**BLEU:** 0%
Referência 5: **Ele chutou a bola** e **perdeu**. **Ele** foi **levado teve os olhos vendados** e **foi mandado para um campo de prisioneiros** por três semanas. **Candidata 5:** **Retrocedeu** a **esfera** e **faltou-a**. Foi **removido blindfolded** e **emitido a um acampamento da prisão** por três semanas.		
Especialista: 0%	**BRAPT:** 86.79%	**BLEU:** 0%
Referência 6: **As** técnicas para **motivar os** jogadores **incluíam** ameaças **de cortar suas pernas** e **jogá-las a** cães **famintos**. **Candidata 6:** Técnicas para **motivate** ameaças **incluídas** jogadores **para eliminar seus pés** e **para jogá-los para** cães **ravenous**.		
Especialista: 0%	**BRAPT:** 83.98%	**BLEU:** 0%

Figura VI.24: Traduções inutilizáveis produzidas pela ferramenta WL (World Lingo).

Nas três sentenças é possível observar palavras não traduzidas. No caso da Candidata 4, por exemplo, esse tipo de erro foi determinante para comprometer a compreensão da sentença. Na candidata 5, traduções equivocadas também contribuíram para que houvesse um desordenamento da referida sentença, tornando-a incompreensível. O mesmo ocorreu em relação à Candidata 6. Os percentuais apresentados pela BRAPT provavelmente se deram pelo fato de haver algumas palavras em comum, ainda que de forma desordenada entre a referência e a candidata. Também fica clara a dificuldade apresentada pelo WL para traduzir palavras como “bola”, “motivar”, “famintos”, dentre outras. Isso indica o nível de fragilidade dessa ferramenta. Cabe ratificar que nestes exemplos, assim como já havia sido indicado anteriormente, o WL apresentou sérias dificuldades com a questão da flexão verbal.

Cabe salientar, que de acordo com as impressões do `especialista avaliador`, a BRAPT também se saiu melhor na maioria das traduções produzidas pelo WL. Houve casos em que a BLEU invalidou sentenças com significados similares às referências e apresentou percentuais mais distantes da avaliação humana na maioria das traduções produzidas pela referida ferramenta. Nos casos em que a BLEU alcançou 100% de compatibilidade, a BRAPT também alcançou, visto que, nesses casos, tratava-se de traduções exatamente iguais às traduções referência. No entanto, houve situações em que a métrica BRAPT indicou certa compatibilidade para sentenças que não apresentavam correspondência mínima com a tradução referência.

Se a métrica BRAPT for utilizada para avaliar a sentença referência “O carro preto pegou fogo.” em comparação com uma candidata que contenha exatamente as mesmas palavras dispostas de forma embaralhada (*e.g.,* “Fogo carro pegou preto o”), certamente essa métrica vai indicar que ambas são compatíveis. Com a métrica BLEU isso já não aconteceria, visto que a alteração na

ordem de uma única palavra já é suficiente para comprometer a compatibilidade BLEU. No exemplo apresentado, em que todas as palavras estão fora de ordem, certamente a compatibilidade BLEU seria 0. No entanto, as principais ferramentas de TAT utilizadas atualmente atingiram um nível de sofisticação que não permite esse tipo de situação. A ausência de termos traduzidos e a falta de ordenação das palavras apresentados nas três sentenças destacadas são incomuns em ferramentas sofisticadas como o BI e o GT. Conclui-se, portanto, que a métrica BRAPT é a mais indicada para esse tipo de ferramenta. No entanto, quando se trata de ferramentas menos sofisticadas, como o WL, essa nova métrica mostra-se menos eficaz e mais suscetível a falhas.

Um fator importante a ser considerado é que ferramentas menos sofisticadas como o World Lingo (WL) tendem a ser aperfeiçoadas. Do contrário, podem sofrer o risco de caírem em desuso, visto que as técnicas utilizadas por ferramentas de TAT atualmente fazem com que as traduções automáticas se aproximem cada vez mais dos resultados produzidos por humanos. Logo, essa dificuldade apresentada pela BRAPT para avaliar algumas traduções produzidas por esse tipo de ferramenta menos sofisticada aparentemente representa um problema de curto ou médio prazo.

Capítulo VII Conclusões

O presente trabalho abordou a questão das Traduções Automáticas de Textos (TATs), os desafios a serem superados por ferramentas que realizam esse tipo de tradução, bem como as métricas existentes para avaliá-las. Os estudos indicaram que essas métricas ainda se mostram ineficazes na tarefa de avaliar as traduções produzidas por ferramentas de TAT muito utilizadas atualmente. Em seguida discute-se sobre o funcionamento das métricas existentes e sobre a importância de adicionar semântica a esse tipo de avaliação. Tomando essa necessidade como ponto de partida, foram apresentadas contribuições que posteriormente foram utilizadas em experimentos que envolveram dez textos jornalísticos escritos originalmente em língua inglesa. Esses textos foram traduzidos por três ferramentas de TAT, que também fizeram parte desse estudo. Todos os experimentos foram realizados sob a supervisão de um humano especialista em traduções.

VII.1 Principais contribuições

Como principal contribuição, foi proposta uma métrica denominada *Bilingual Rating of Psycholinguistic Perspectives in Translations* (BRAPT) como alternativa capaz de avaliar TATs considerando aspectos semânticos presentes nas traduções. Esse tipo de avaliação foi possível com a utilização de um algoritmo denominado `calcBRAPT` e da uma ferramenta de análise textual denominada LIWC. Essa ferramenta possui um léxico capaz de classificar palavras em categorias que refletem aspectos linguísticos e psicológicos. Cada palavra presente no LIWC pode estar relacionada a diversas categorias que refletem aspectos linguísticos e/ou psicológicos e cada categoria contém diversas palavras associadas a ela. Assim, cada palavra identificada em uma sentença pode ser contabilizada em diversas categorias do LIWC. Com o auxílio desse léxico, sentenças referência e candidata puderam ser transformadas em vetores de categorias psicolinguísticas. A compatibilidade entre essas duas sentenças (*i.e.,* esses dois vetores) pôde ser verificada com a utilização da medida de similaridade do cosseno.

Com o auxílio do LIWC foi possível propor um outro algoritmo denominado `calcDPC`. Esse algoritmo proporcionou uma análise individual de cada categoria psicolinguística (*i.e.,* de cada aspecto linguístico ou psicológico) presente em uma sentença, seja ela a referência ou a can-

didata. Por meio dessa análise foi possível verificar as divergências de determinado aspecto linguístico (*e.g.,* verbos, advérbios, preposições) ou psicológico (*e.g.,* raiva, ansiedade, emoções positivas, aspectos cognitivos) no processo de tradução.

Por meio da métrica proposta e do algoritmo `calcDPC` também foi possível avaliar três ferramentas de TAT muito utilizadas em ambiente *web*: google Tradutor (GT), Bing Tradutor (BI) e World Lingo (WL). Os resultados foram considerados bem consistentes com os já apresentados em estudos anteriores e com as análises de um especialista em traduções (o `especialista avaliador`) que atuou na supervisão de todos os testes, incluindo a comparação das métricas.

Algumas das contribuições se encontram desmembradas nas seguintes publicações realizadas durante os estudos que compreendem o presente trabalho:

- RODRIGUES, Rafael Guimarães et al. TATMaster: Psycholinguistic Divergences in Automatically Translated Texts. In: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web. ACM, 2017. p. 205-208.

- RODRIGUES, Rafael Guimarães; GUEDES, Gustavo Paiva. 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. ACM, 2017. p. 161-164.

- RODRIGUES, Rafael Guimarães et al. *A new metric for translation evaluation based on psycholinguistic perspectives*. Artigo submetido, em 04/04/2018, ao periódico IEEE Latin American Transactions (aguardando resultado).

VII.2 Experimentos realizados

Neste trabalho foi apresentada uma avaliação experimental aplicando o algoritmo `calcDPC` a cada sentença dos dez textos utilizados nesse estudo. Essa avaliação indicou, por exemplo, que a ferramenta WL apresenta desempenho muito ruim em relação a flexão de verbos e em relação a aspectos psicológicos. Da mesma forma também foi possível verificar que a ferramenta GT se comporta muito bem com verbos auxiliares e em relação aos aspectos relativos a raiva. Esse tipo de informação pode ser útil para o desenvolvimento de novas métricas, para o aprimoramento das métricas existentes ou até mesmo em estudos que visam melhorar a qualidade das ferramentas utilizadas para a realização de TATs.

Os experimentos também consistiram em comparar a métrica proposta (BRAPT) com o atual estado da arte (*i.e.,* a métrica BLEU). Essa métrica, por ser baseada no pareamento exato e ordenado de palavras, mostrou-se limitada, incapaz de avaliar a semântica presente nas sentenças referência e candidata e, consequentemente, incapaz de considerar aspectos importantes das traduções, como a utilização de sinônimos, a inversão de ordem de palavras. Esses aspectos não

deveriam representar taxas consideráveis de incompatibilidade em métricas que avaliam a qualidade de TATs, visto que, na maioria dos casos, não representam comprometimento significativo das traduções. De acordo com os dados apresentados e com as impressões do `especialista avaliador`, a métrica BRAPT mostrou-se mais adequada do que a métrica BLEU pela sua capacidade de adicionar semântica à avaliação, considerando aspectos linguísticos e psicológicos presentes nas sentenças envolvidas no processo, oferecendo, portanto, a possibilidade de realizar uma avaliação muito mais flexível e menos suscetível a invalidar traduções candidatas similares às traduções referências.

Em relação às três ferramentas de TAT estudadas, as métricas utilizadas (BLEU e BRAPT), estudos anteriores e as avaliações do especialista indicaram que as ferramentas BI e GT apresentam um desempenho bem semelhante e traduções muito mais próximas das traduções referência do que a ferramenta WL, que mostrou-se bem menos sofisticada. Os testes com o algoritmo `calcDPC` também indicaram a fragilidade das traduções produzidas por essa ferramenta sob diversos aspectos linguísticos e psicológicos, com destaque para a dificuldade para traduzir palavras traduzidas facilmente pelo BI e pelo GT e para a dificuldade com relação à flexão verbal.

VII.3 Limitações

Durante os experimentos foram detectadas algumas limitações em relação ao LIWC e também, em alguns casos, em relação à utilização da métrica BRAPT. Os testes mostraram que a versão do LIWC para o português do Brasil ainda não conta com palavras relevantes com “escritora”, “focalizar”, “inabitável”, “legalização”, “acessar”, “retroceder”, dentre outras. Em relação à BRAPT, em alguns casos, essa nova métrica apresentou problemas para avaliar traduções muito ruins produzidas por ferramentas de TAT pouco sofisticadas, como o World Lingo, por exemplo. Com relação a esses problemas (*i.e.,* limitações), entende-se que a tendência é a de que eles sejam resolvidos em médio ou curto prazo, visto que traduções muito ruins são incomuns nas ferramentas de TAT mais sofisticadas e que as ferramentas menos sofisticadas tendem a evoluir ou a caírem em desuso. Quanto ao LIWC, há estudos em andamento para o lançamento de novas versões, o que certamente representará uma melhora significativa em relação às palavras não encontradas.

VII.4 Considerações finais

Os experimentos foram considerados relevantes, visto que apresentaram resultados consistentes com os que já existiam na literatura e com as análises de um profissional especializado em traduções de textos. A métrica BRAPT se mostrou superior ao estado da arte, especialmente quando utilizada para avaliar de TAT que produzem bons resultados. Superioridade essa,

atestada também pelo especialista em traduções (*i.e.,* o `especialista avaliador`).

Durante os experimentos também foi possível avaliar a qualidade de três ferramentas de TAT utilizadas em ambiente *web*. As avaliações realizadas pela BRAPT acerca da qualidade dessas ferramentas também coincidiram com as avaliações do especialista. Em trabalhos futuros pretende-se utilizar a métrica BRAPT com outras versões do LIWC ou até mesmo com outros léxicos afetivos similares, além de utilizar outras técnicas capazes de verificar a similaridade entre sentenças, conhecidas na literatura ao invés da similaridade do cosseno.

Por fim, entende-se que este estudo vem preencher uma lacuna no que diz respeito à adição de semântica (*e.g.,* consideração de aspectos linguísticos e psicológicos) ao processo de avaliação de TATs. As novas perspectivas apresentadas no presente trabalho representam uma contribuição relevante e indicam uma nova direção a ser seguida em outros estudos que envolvam traduções automáticas de textos e métricas propostas para avaliá-las. Outras contribuições presentes neste trabalho também se mostraram relevantes. Por meio do algoritmo `calcDPC`, foi apresentada a possibilidade de identificar transformações relevantes e reveladoras sobre o funcionamento de ferramentas de TAT ocorridas em cada aspecto linguístico ou psicológico de uma sentença proveniente desse tipo de tradução.

Referências Bibliográficas

Aly, A. and Tapus, A. (2013). A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 325–332. IEEE Press.

Baskaya, O., Yildiz, E., Tunaoglu, D., Eren, M. T., and Doğruöz, A. S. (2017). Integrating meaning into quality evaluation of machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 210–219.

Beck, D. E. (2009). Aprimorando o tratamento de expressões multipalavras em um tradutor automatico baseado em regras.

Brandão, H. H. N. (2009). Representações da escrita: estereotipia e singularidade enunciativa. *Scripta*, 13(24):111–128.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluation the role of bleu in machine translation research. In *EACL*, volume 6, pages 249–256.

Cer, D., Manning, C. D., and Jurafsky, D. (2010). The best lexical metric for phrase-based statistical mt system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 555–563. Association for Computational Linguistics.

Chang, P.-C., Galley, M., and Manning, C. D. (2008). Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232. Association for Computational Linguistics.

da Ponte Junior, L. A., Guedes, G. P., Bezerra, E., and Maracana, A. (2016). Inferindo o sexo de usuarios de redes sociais utilizando o liwc em português do brasil.

da Rocha, R. L. d. A. (2007). Adaptive technology applied to natural language processing. *IEEE Latin America Transactions*, 5(7):544–551.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, 13:1–10.

de Melo, F. R., de Oliveira Matos, H. C., and Dias, E. R. B. (2015). 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*, 5(3):33–45.

de Oliveira, M. G., Oliveira, E., and Marchesi, R. Z. (2009). Um qasystem para interação de alunos em avaliações somativas a distância. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 1.

Di Thommazo, A., Malimpensa, G., de Oliveira, T. R., Olivatto, G., and Fabbri, S. C. (2012). Requirements traceability matrix: Automatic generation and visualization. In *Software Engineering (SBES), 2012 26th Brazilian Symposium on*, pages 101–110. IEEE.

D'Mello, S. K. and Graesser, A. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5(4):304–317.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

dos Santos Machado, A. (1995). *Matemática: temas e metas*. Atual.

Dzindolet, M. T. and Pierce, L. G. (2005). Using a linguistic analysis tool to detect deception. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 563–567. SAGE Publications Sage CA: Los Angeles, CA.

Feldman, R. and Dagan, I. (1995). Knowledge discovery in textual databases (kdt). In *KDD*, volume 95, pages 112–117.

Ferreira, A. A. (2012). Investigando o processamento cognitivo de tradutores profissionais em tradução direta e inversa no par linguístico inglês-português. *Cadernos de traduçao*, 1(29):73–92.

Filho, Pedro P.Balage;Pardo, T. A. S. R. M. A. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Finch, A. M., Akiba, Y., and Sumita, E. (2004). How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? In *LREC*.

França, C. L., Weizenmann da Matta, K., and Dornelles Alves, E. (2012). Psicologia e educação a distância: uma revisão bibliográfica. *Psicologia Ciência e Profissão*, 32(1).

Gelbukh, A., Sidorov, G., Han, S.-Y., and Hernández-Rubio, E. (2004). Automatic enrichment of very large dictionary of word combinations on the basis of dependency formalism. In *Mexican International Conference on Artificial Intelligence*, pages 430–437. Springer.

Golbeck, J., Robles, C., and Turner, K. (2011). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*, pages 253–262. ACM.

Guedes, G. P., Bezerra, E., Ferrari, L., and Duarte, F. (2016). Gender differences in the use of portuguese in social networks: Evidence from liwc. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 339–342. ACM.

Guimarães, R. (2016). Croca-cromoterapia e computação afetiva: auxiliando os. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, volume 13069, page 901. ACM.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62.

JOHASSON, S. (2004). The individual and the species in the cultural evolution of language. In *EELC, Brussels*.

Jones, D., Shen, W., and Herzog, M. (2009). Machine translation for government applications. *Lincoln Laboratory Journal*, 18(1).

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Li, H., Graesser, A. C., and Cai, Z. (2014). Comparison of google translation with human translation. In *FLAIRS Conference*.

Liddy, E. D. (2001). Natural language processing. In *Syracuse University SURFACE*.

Linares, J. A. G. (2005). Rich linguistic knowledge for empirical machine translation.

Mauser, A., Hasan, S., and Ney, H. (2008). Automatic evaluation measures for statistical machine translation system optimization. In *LREC*.

Miller, G. A. (1991). The science of words. Scientific American Library.

Morris, A. C., Maier, V., and Green, P. (2004). From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of liwc2007.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001a). *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001b). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pereira, J. W., Gonçalves, M. R. B., and Santos, M. T. P. (2013). Pré-processamento para recuperação de informação em textos históricos do século xix. In *Proceedings of the Symposium on Knowledge Discovery, Mining and Learning. Sao Carlos, SP, Brazil*.

Picard, R. W. (1995). Affective computing. *The MIT Press, Cambridge (MA)*.

Picard, R. W. (1997). Affective computing. *The MIT Press, Cambridge (MA)*, 167:170.

Rezende, S., Pugliesi, J., Melanda, E., and de Paula, M. (2003). Mineração de dados, chapter 12. *Rezende (2003)*, 2:1–6.

Rodrigues, R. G., Gomes, R. R., Rodrigues, K. T., and Guedes, G. P. (2017a). Tatmaster: Psycholinguistic divergences in automatically translated texts. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, pages 205–208. ACM.

Rodrigues, R. G. and Guedes, G. P. (2017). 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*, pages 161–164.

Rodrigues, R. G., Paiva Guedes, G., and Ogasawara, E. (2016). Towards a model for personality-based agents for emotional responses. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 359–362. ACM.

Rodrigues, R. G., Pereira, W. W., Bezerra, E., Guedes, G. P., Maracana, A., and de Janeiro-RJ-Brasil, R. (2017b). Inferência de idade utilizando o liwc: identificando potenciais predadores sexuais. In *XXXVII Congresso da Sociedade Brasileira de Computaç ao*.

Sales, S. G. (2011). Tradução automática: os processos da tradução mediada por computador. *Saberes em perspectiva*, 1(1):19–37.

Santiago, M. L. (2013). Avaliação da tradução automática: Bulas versus outros géneros textuais.

Santos, J. T. L., Anastacio, I. M., and Martins, B. E. (2015). Named entity disambiguation over texts written in the portuguese or spanish languages. *IEEE Latin America Transactions*, 13(3):856–862.

Schardong, G. G., Silva, L. J., Winck, A. T., and Pozzer, C. T. (2013). Agrupamento de dados baseado em mean shift aplicado a legendas de séries televisivas. In *Proceedings of the Symposium on Knowledge Discovery, Mining and Learning. Sao Carlos, SP, Brazil*.

Tatai, G. and Laufer, L. (2004). Extraction of affective components from texts and their use in natural language dialogue systems. *Acta Cybern.*, 16(4):625–642.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Tavares, R., Guedes, G. P., da Fonseca, C. S., and de Janeiro-RJ-Brasil, R. (2017). Classificaç ao de filmes: uma abordagem utilizando o liwc. In *XXXVII Congresso da Sociedade Brasileira de Computaç ao*.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *Eurospeech*.

Weininger, M. J. (2004). Tm & mt na tradução técnica globalizada–tendências e conseqüências. *Cadernos de tradução*, 2(14):243–263.

White, J., O'Connell, T., and O'Mara, F. (1994). The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*, pages 193–205.

Zeng, X., Chao, L. S., Wong, D. F., Trancoso, I., and Tian, L. (2014). Toward better chinese word segmentation for smt via bilingual constraints. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1360–1369.

Capítulo VIII Anexos

VIII.1 Representatividade de cada aspecto psicolinguístico nos textos utilizados

Tabela VIII.1: Representatividade de aspectos psicolinguísticos na tradução referência

Categoria	Representatividade
1-funct	11.38%
131-cogmech	8.89%
250-relativ	5.65%
500-nfound	4.74%
121-social	3.77%
17-preps	3.67%
11-verb	3.49%
138-incl	2.91%
2-pronoun	2.80%
253-time	2.65%
135-tentat	2.49%
150-ingest	2.42%
252-space	2.25%
18-conj	2.05%
14-present	1.91%
9-ipron	1.83%
12-auxverb	1.74%
3-ppron	1.69%
125-affect	1.67%
140-percept	1.63%
251-motion	1.59%
146-bio	1.49%
139-excl	1.48%
124-humans	1.45%
20-quant	1.42%
10-article	1.40%
22-swear	1.32%
137-inhib	1.21%
134-discrep	1.21%
132-insight	1.20%
6-you	1.15%
7-shehe	1.11%
133-cause	1.11%

Tabela VIII.2: Representatividade de aspectos psicolinguísticos na tradução referência

Categoria	**Representatividade**
355-achieve	1.07%
16-adverb	0.90%
13-past	0.88%
126-posemo	0.78%
143-feel	0.77%
127-negemo	0.75%
8-they	0.65%
136-certain	0.64%
147-body	0.60%
21-number	0.58%
358-money	0.58%
354-work	0.51%
141-see	0.50%
148-health	0.47%
130-sad	0.41%
129-anger	0.40%
356-leisure	0.37%
360-death	0.30%
142-hear	0.25%
462-assent	0.25%
19-negate	0.23%
149-sexual	0.22%
357-home	0.18%
15-future	0.17%
123-friend	0.17%
359-relig	0.14%
122-family	0.12%
128-anx	0.10%
4-i	0.09%
5-we	0.08%
463-nonfl	0.03%
464-filler	0.00%

VIII.2 Divergências de aspectos psicolinguístico nas TATs

Tabela VIII.3: Médias de divergências de aspectos psicolinguísticos verificados em TATs

Categoria (Aspecto)	Tradução GT	Tradução BI	Tradução WL
1-funct	3.15%	1.81%	3.97%
10-article	8.50%	5.41%	20.45%
11-verb	-0.29%	0.57%	-5.16%
12-auxverb	0.00%	8.90%	5.95%
121-social	1.82%	0.26%	-2.12%
122-family	-16.67%	-33.33%	-41.67%
123-friend	-5.88%	-11.76%	-17.65%
124-humans	-4.14%	2.68%	-7.59%
125-affect	-5.99%	-6.59%	-20.36%
126-posemo	6.02%	8.24%	6.02%
127-negemo	-16.00%	-20.00%	-44.00%
128-anx	-40.00%	-20.00%	-50.00%
129-anger	2.44%	-10.00%	-42.50%
13-past	2.22%	2.22%	-9.09%
130-sad	-17.07%	-21.95%	-58.54%
131-cogmech	-0.90%	-0.34%	-0.79%
132-insight	1.64%	7.69%	-5.00%
133-cause	-7.21%	-10.81%	-12.61%
134-discrep	-2.48%	-3.31%	-12.40%
135-tentat	-3.61%	-3.61%	-2.01%
136-certain	-3.12%	-4.69%	16.88%
137-inhib	6.20%	6.20%	2.42%
138-incl	-2.75%	2.35%	-5.84%
139-excl	5.13%	6.33%	6.33%
14-present	6.83%	10.75%	3.05%
140-percept	-12.27%	-18.40%	-15.34%
141-see	-14.00%	-18.00%	-8.00%
142-hear	-8.00%	-8.00%	0.00%
143-feel	-15.58%	-20.78%	-22.08%
146-bio	-2.01%	10.24%	-9.40%
147-body	-8.33%	6.25%	-3.33%
148-health	-6.38%	2.08%	-31.91%
149-sexual	-4.55%	0.00%	18.52%
15-future	29.17%	19.05%	19.05%
150-ingest	-4.96%	-13.64%	-9.50%

Tabela VIII.4: Médias de divergências de aspectos psicolinguísticos verificados em TATs

Categoria (Aspecto)	Tradução GT	Tradução BI	Tradução WL
16-adverb	-4.44%	-2.22%	15.09%
17-preps	-1.36%	-2.45%	4.68%
18-conj	0.00%	2.84%	0.49%
19-negate	17.86%	4.17%	0.00%
2-pronoun	8.79%	6.35%	2.78%
20-quant	8.39%	6.58%	16.47%
21-number	1.69%	-3.45%	10.77%
22-swear	-6.06%	-12.12%	0.00%
250-relativ	0.53%	0.18%	-0.35%
251-motion	-15.09%	-10.69%	-22.01%
252-space	6.64%	2.60%	10.00%
253-time	-3.02%	-7.92%	-4.15%
3-ppron	7.14%	2.87%	3.43%
354-work	12.07%	-7.84%	19.05%
355-achieve	-6.54%	-9.35%	-9.35%
356-leisure	7.50%	7.50%	-13.51%
357-home	-5.56%	-16.67%	-11.11%
358-money	-12.07%	7.94%	-3.45%
359-relig	-21.43%	-28.57%	-57.14%
360-death	-3.33%	-6.67%	-3.33%
4-i	0.00%	0.00%	0.00%
462-assent	7.41%	13.79%	3.85%
463-nonfl	0.00%	0.00%	0.00%
464-filler	0.00%	0.00%	0.00%
5-we	0.00%	0.00%	33.33%
500-nfound	-3.80%	3.46%	21.13%
6-you	7.26%	7.26%	19.58%
7-shehe	4.31%	5.93%	-13.51%
8-they	15.58%	12.16%	9.72%
9-ipron	8.50%	8.04%	-1.09%

VIII.3 Outros exemplos de perdas de aspectos psicolinguísticos em traduções

Tabela VIII.5: Perdas na categoria `raiva` com o BI (Texto 1, sentença 12).

Sentença	**Perda de 38.96% na categoria `129-anger` com o BI**	**fp**	**rep**
Refer.	Além de **apresentar maior duração** a verdadeira depressão clínica é também mais intensa do que um **simples** caso de **tristeza.**	2	1.72%
Cand. BI	Além de **ser mais duradouro** a verdadeira depressão clínica também é mais intensa do que um caso de **blues.**	1	1.05%

Tabela VIII.6: Perdas na categoria `aspectos familiares` com o BI (Texto 2, sentença 8).

Sentença	**Perda de 44.86% na categoria `122-family` com o BI**	**fp**	**rep**
Refer.	Ao invés de se **sentirem desolados sem os filhos** muitos casais como os Colliers se **acham envolvidos** no que é comumente conhecido como segunda lua de mel **após os** filhos **deixarem o lar.**	2	1.07%
Cand. BI	Ao invés de se **sentir desprovido de crianças** muitos casais como os Colliers se **encontram varrido** no que é comumente conhecido como **uma** segunda lua de mel **depois de seus** filhos **sair.**	1	0.59%

Tabela VIII.7: Perdas na categoria `3ª pessoa singular` com o WL (Texto 6, sentença 15).

Sentença	**Perda de 67.86% na categoria `7-shehe` com o WL**	**fp**	**rep**
Refer.	**No momento em que retorna ela** sabe **pelo** seu **relógio de bordo** que 20 anos **se** passaram. **Ela** agora tem 45 anos.	4	3.48%
Cand. WL	**Pelo tempo onde a aterra** sabe **de** seu **pulso de disparo on-board** que 20 anos passaram. Tem agora 45 anos **velha.**	1	1.24%