TATModel: Em Direção a um Novo Modelo para Avaliação de Traduções Automáticas de Texto

R. G. Rodrigues[1], G. P. Guedes[1]

CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Brazil
{rafael.rodrigues, gustavo.guedes}@cefet-rj.br

Abstract. This work aims to propose a new model capable of identifying and quantifying psycholinguistic changes in the translation of texts from English to Brazilian Portuguese. This model uses a textual analysis tool named LIWC to classify words into psychological and linguistic categories. The word count in each category is used to identify psychological and linguistic changes in the translanted texts. The experiments indicate promising results.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining; I.2.7 [**Natural Language Processing**]: Text analysis; I.7.2 [**Document Preparation**]: Standards; J.5 [**ARTS AND HUMANITIES**]: Language translation

Keywords: automatic translation, psycholinguistic changes, textual analysis

1. INTRODUÇÃO

A Tradução Automática de Textos (TAT) consiste em traduzir palavras e expressões de um idioma para outro. As ferramentas de TAT são bastante utilizadas nos dias atuais e atuam como importantes facilitadores da comunicação na era da globalização. Essas ferramentas fazem uso de algumas técnicas como, por exemplo, as Memórias de Tradução (MT). Essa técnica, utilizada por algumas ferramentas, consiste em armazenar e utilizar dados de traduções anteriores a serem utilizados de acordo com o reaparecimento de elementos idênticos ou muito parecidos, configurando economia significativa de tempo no processo de tradução [Weininger, 2004]. As ferramentas de TAT, no entanto, ainda carecem de revisão por parte de um profissional especializado. Esse tipo de trabalho depende de revisão humana por conta de aspectos linguísticos e de conhecimentos característicos do ser humano, cuja interferência ainda é indispensável para garantir a confiabilidade da tradução [Sales, 2011].

A tradução de palavras ambíguas (*e.g., blue*, *interest*, *rare*), expressões idiomáticas e até mesmo expressões simples refletem alguns dos desafios ainda a serem superados pelas ferramentas de TAT. Nesse estudo, utilizamos três ferramentas muito conhecidas que apresentaram dificuldades para realizar traduções da língua inglesa para o português do Brasil, quando avaliadas de acordo com aspectos linguísticos, psicológicos e até mesmo com o contexto das palavras ou expressões a serem traduzidas.

O objetivo deste trabalho é propor um modelo capaz de quantificar estatisticamente as divergências psicolinguísticas ocorridas nas traduções realizadas por ferramentas de TAT, quando comparadas com traduções realizadas por um especialista em traduções. A partir deste ponto, utilizaremos o termo **tradução candidata** para as traduções realizadas por ferramentas de TAT e **tradução referência** para as traduções realizadas por um especialista.

As mudanças psicolinguísticas ocorridas na comparação entre uma **tradução candidata** e uma **tradução referência** são identificadas por uma ferramenta de mineração de textos denominada

Linguistic Inquiry Word and Count (LIWC) [Pennebaker et al., 2001]. Essa ferramenta é capaz de quantificar e classificar palavras em categorias linguísticas e psicológicas.

O restante desse trabalho está organizado da seguinte forma: na Seção 2, discorremos sobre a TAT, o LIWC e trabalhos relacionados. Na Seção 3, descrevemos nosso modelo. Na seção 4 são apresentados os resultados experimentais e na Seção 5, apresentamos as considerações finais e contribuições.

2. A TRADUÇÃO AUTOMÁTICA DE TEXTOS E O LIWC

A TAT tem o objetivo de realizar traduções de um idioma para outro, buscando manter a equivalência com o texto ou mensagem original. A TAT surgiu nos anos 50 e, assim como boa parte das tecnologias, foi motivada por questões militares. Atualmente, a maioria dessas ferramentas funciona em ambiente *web*. O estudo realizado em [de Melo et al., 2015] indica que ainda há muito a se fazer nessa área e aborda uma métrica denominada BLEU (*Bilingual Evaluation Understudy*). Essa métrica consiste em comparar uma **tradução candidata** com uma **tradução referência**, com o objetivo de avaliar a equivalência entre essas traduções. No entanto, os autores alertam que a BLEU resume-se a apenas uma dentre diversas formas quantitativas de avaliar a precisão das traduções e que, sozinha, não é suficiente para avaliar a qualidade de uma TAT.

A confiabilidade de uma TAT ainda depende de revisão humana, portanto a possibilidade de extrair características humanas dos textos representa um ganho considerável nesse processo. Nesse sentido, há estudos que buscam fazê-lo por meio da mineração de textos [Pereira et al., 2013, Schardong et al., 2013]. Alguns desses estudos utilizam o LIWC para essa finalidade [Araújo et al., 2013, Rodrigues et al., 2017]. O léxico contido no LIWC categoriza palavras em uma ou mais categorias, que refletem aspectos linguísticos (*e.g., ppron, adverb*), psicológicos (*e.g., negemo, anger*) e sociais (*e.g., family, human*). Em sua versão para o português do Brasil, o LIWC possui um léxico com 127.149 palavras distribuídas em 64 categorias (27 categorias principais e 37 subcategorias) [Balage Filho et al., 2013].

Os estudos para a criação do LIWC envolveram profissionais de diversas áreas e citaram a existência de duas categorias amplas de palavras com propriedades psicométricas e psicológicas: as **palavras de conteúdo** (*e.g.,* substantivos, verbos regulares, adjetivos e advérbios) transmitem o conteúdo da comunicação; as **palavras de estilo** (*e.g.,* pronomes, preposições, artigos, conjunções, verbos auxiliares) revelam o estilo de comunicação dos indivíduos [Tausczik and Pennebaker, 2010]. Embora haja muito mais palavras de conteúdo do que de estilo, essas últimas representam a maior parte de uma comunicação escrita ou falada [Tausczik and Pennebaker, 2010].

3. O MODELO TATMODEL

O modelo proposto neste trabalho baseia-se na alocação de categorias em vetores gerados para representar cada sentença de um texto t. Cada vetor $\vec{v}$ contém n posições correspondentes às categorias de palavras utilizadas pelo TATModel. Além das 27 categorias principais do LIWC, foi adicionada uma categoria extra para alocar as palavras não existentes nessa ferramenta. Dessa forma, o número de categorias do TATModel foi definido por $n = 28$. Com isso, um texto com 10 sentenças, por exemplo, é representado por 10 vetores $\vec{v}$, cada um com 28 posições.

Cada posição x_i de um vetor $\vec{v}$ representa uma categoria. Para fins de computação do valor de $\vec{v}$, cada palavra pertencente à sentença s foi identificada no TATModel e retornou um conjunto de categorias. As posições x_i referentes a cada uma das categorias retornadas foram incrementadas em $\vec{v}$. Em seguida, calculou-se o percentual de representatividade de cada categoria x_i em relação a s. A Figura 1 ilustra o vetor resultante de uma sentença em que 9.75% das palavras se enquadraram na categoria representada por x_2 e 7.89% das palavras se enquadraram na categoria representada por x_6. O comprimento de $\vec{v}$, como já foi mencionado, é dado por n.

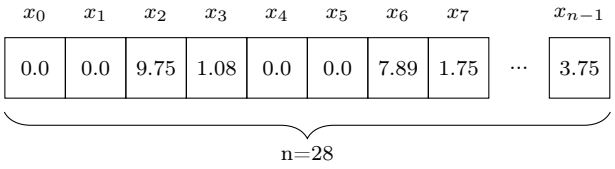

Fig. 1. Vetor $\vec{v}$ ilustrando o percentual de representatividade de cada categoria x_i em relação a uma sentença s.

O referido modelo consiste em analisar uma **tradução referência** comparada a uma **tradução candidata**. A análise dos textos é feita sentença a sentença seguindo os seguintes passos:

(1) Gerar dois vetores: **vetor referência** e **vetor candidato**, correspondentes, respectivamente, às sentenças provenientes da **tradução referência** e da **tradução candidata**.
(2) Alocar, em cada posição do vetor, a contagem de palavras para sua respectiva categoria.
(3) Atualizar os vetores com os percentuais de representatividade de cada posição em relação ao somatório dos valores de todas as posições de cada vetor, que sempre será 100.0.
(4) Gerar um **vetor de mudanças** psicolinguísticas a partir do módulo da subtração entre os vetores **referência** e **candidato**. Por representar os outros dois vetores citados, cada um com somatório igual a 100.0, o valor contido em cada posição será dividido por 2 a fim de preservar a proporção de representatividade de cada categoria. O somatório desse novo vetor varia entre 0.0 e 100.0.

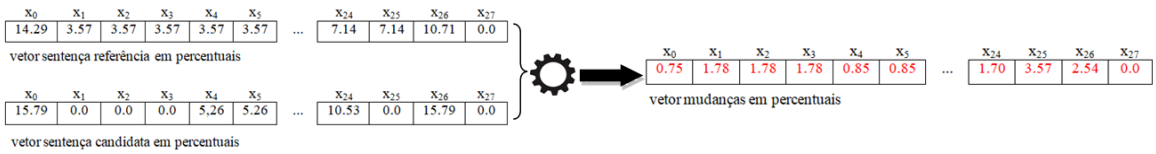

Fig. 2. Diferença entre vetor referência e vetor candidato, representando as mudanças identificadas entre duas sentenças.

Ao observar a figura 2, percebemos que o **vetor mudança** apresentou mudanças em praticamente todas as categorias ilustradas. Apenas a categoria x_{27} permaneceu com valor 0.0, por não apresentar divergências entre os valores contidos na posição x_{27} dos vetores referência e candidato.

O **vetor mudança** sempre vai conter valores entre 0.0 e 100.0, pois esse vetor guarda o módulo da subtração dos outros dois vetores. O valor 0.0 indica que a compatibilidade foi total e o valor 100.0 indica que houve incompatibilidade total entre a **tradução referência** e a **tradução candidata**.

4. RESULTADOS EXPERIMENTAIS

A fim de testar o modelo proposto, selecionamos oito textos traduzidos por um profissional e por três ferramentas de TAT conhecidas: Google Tradutor (GT), Bing Tradutor (BT) e WorldLingo (WL)[1]. O WL apresentou o pior desempenho em todos os textos. Os resultados são mostrados na tabela 1.

Table I. Mudanças de características psicolinguísticas em ferramentas de TAT comparadas com a tradução referência.

	Tex1	Tex2	Tex3	Tex4	Tex5	Tex6	Tex7	Tex8	**Média**
Google	**14.07**%	12.09%	**8.91**%	**12.45**%	**8.13**%	**9.67**%	12.93%	**12.35**%	**11.33**%
Bing	14.33%	**12.02**%	9.29%	12.68%	9.53%	10.96%	**11.58**%	12.45%	11.61%
WorldLingo	21.15%	14.60%	15.88%	18.07%	15.67%	13.28%	15.20%	18.47%	16.54%

[1]https://www.google.com.br/translator, https://www.bing.com/translator e http://www.worldlingo.com/

Podemos observar, **em negrito**, os menores percentuais de mudanças em comparação com o texto referência. O menor percentual (8.13%) foi obtido na tradução do texto 7 pelo GT, que obteve os melhores resultados em outros cinco textos e na média geral. Cabe lembrar que quanto menor for o percentual de mudanças, mais próxima estará a **tradução candidata** da **tradução referência**.

Ao analisar as sentenças de todos os textos em conjunto, ou seja, 123 sentenças, as mudanças psicolinguísticas variaram de 0.0% (GT e BT) a 48.79% (WL). Apesar do pior desempenho da ferramenta WL, em algumas sentenças as ferramentas GT e BT também chegaram a apresentar percentuais de mudanças significativos, em torno de 41.29% (GT) e 47.63% (BT). Assim, fica claro que a ausência de interpretação humana na tradução ainda é um complicador na TAT.

5. CONCLUSÕES E CONTRIBUIÇÕES

O objetivo desse estudo foi a proposição de um modelo capaz de identificar mudanças psicolinguísticas ocasionadas na TAT em comparação com traduções referência, realizadas por um especialista. Para isso, foi utilizada a ferramenta LIWC, que se mostrou eficaz na tarefa de identificar e quantificar essas mudanças. Consideramos os resultados significantes e entendemos que o modelo se mostrou eficiente.

As ferramentas de TAT precisam evoluir no tocante a determinados aspectos que ainda dependem de intervenção humana. Durante o estudo não foram encontrados trabalhos que utilizam o LIWC em português do Brasil para identificar as mudanças psicolinguísticas ocorridas na TAT. Dessa forma, consideramos que esse trabalho preenche uma lacuna em uma área que ainda carece de novos estudos e pode contribuir para novos trabalhos que envolvam a tradução automática de textos.

Referências

Araújo, M., Gonçalves, P., and Benevenuto, F. Measuring sentiments in online social networks. In *Proceedings of the 19th Brazilian symposium on Multimedia and the web*. ACM, pp. 97–104, 2013.

Balage Filho, P. P., Pardo, T. A., and Aluisio, S. M. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*. pp. 215–219, 2013.

de Melo, F. R., de Oliveira Matos, H. C., and Dias, E. R. B. 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, 2015.

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

Pereira, J. W., Gonçalves, M. R. B., and Santos, M. T. P. 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*, 2013.

Rodrigues, R. G., Pereira, W. W., Bezerra, E., and Guedes, G. P. Inferência de idade utilizando o liwc: identificando potenciais predadores sexuais. In *Proceedings of the 6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2017.

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

Schardong, G. G., Silva, L. J., Winck, A. T., and Pozzer, C. T. 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*, 2013.

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

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