

A Comparative Study of Bing Translator and Google Translate Accuracy Using Automatic Evaluation in a English-Portuguese *Corpus*

Emillie Rebecca Bastos Dias (Universidade Estadual de Goiás, Goiás, Brasil) - emillie.info@gmail.com

Hellen C. de Oliveira Matos (Universidade Estadual de Goiás, Goiás, Brasil) - hellen.sistemas@gmail.com

Francisco Ramos de Melo - (Universidade Estadual de Goiás, Goiás, Brasil) - francisco.melo@ueg.br

Automatic translations are not completely reliable. Even with the progress achieved by the current automatic translators, a human review is recommended to ensure the quality of the outputs. It is possible to estimate the quality of automatic translations through evaluations. The results can help identify errors and also assist users in the choice of which translator fits better for their own needs.

This paper gives a comparison between Bing Translator and Google Translate accuracy. The BLEU score is a widely used method to evaluate machine translations using human reference translations. The texts used in this paper comprise the corpus English-Portuguese that was constructed for this purpose. The corpus will be evaluated by using the BLEU metric implemented by the MT-Eval script. Finally, the scores will be provided for comparative purposes.

Keywords: Machine Translation Evaluation, Bing Translator, Google Translate, BLEU.

Estudo Comparativo da Precisão dos Tradutores Bing e Google Tradutor por meio de Avaliação Automática em um *Corpus* Linguístico inglês-português

As traduções automáticas não são totalmente confiáveis. Mesmo com o progresso já alcançado pelos atuais tradutores automáticos, ainda recomenda-se que haja uma revisão humana após a tradução para que sejam corrigidas algumas falhas.

É possível estimar a qualidade das traduções automáticas por meio de avaliações. Os resultados obtidos permitem que sejam identificadas falhas com a finalidade de aprimorar técnicas e ferramentas de tradução automática ou até mesmo auxiliar usuários na escolha de qual tradutor automático utilizar.

Este artigo apresenta e discute os resultados obtidos em uma avaliação automática comparativa entre os tradutores online Bing Tradutor e o Google Tradutor. A métrica BLEU é amplamente adotada para avaliar a precisão de traduções automáticas utilizando traduções humanas como referência. Por esse motivo, a BLEU foi implementada por meio do script MT-Eval em um corpus linguístico inglês-português, que foi construído para a finalidade desse estudo. Ao final, as pontuações resultantes da avaliação entre os tradutores selecionados são apresentadas de forma comparativa.

Palavras-chave: Avaliação de Tradutores Automática, Bing Tradutor, Google Tradutor, BLEU.

1. Introdução

A tradução automática é o processo de utilização da computação para converter uma mensagem de uma linguagem natural à outra com a preservação da equivalência do conteúdo da mensagem. Os primeiros projetos nessa área foram desenvolvidos com motivações militares durante a Guerra Fria (Hutchins & Sommers, 1992) e, desde então, a tradução automática tornou-se um meio facilitador de relações sócio-econômicas, científicas, religiosas, e vários outros propósitos de comunicação advindos da globalização. Em consequência, várias ferramentas de tradução automática foram desenvolvidas, como o Bing Tradutor e o Google Tradutor, amplamente utilizados na atualidade.

Embora seja um campo de estudo originado há mais de sessenta anos, a tradução automática é uma área pouco desenvolvida e explorada. As propostas para solucionar a insatisfação com as traduções mecanizadas tiveram poucos avanços desde os primórdios de seus estudos. A qualidade desse tipo de tradução ainda depende da revisão humana para garantir um texto confiável. A aplicação da tradução automática pode se tornar inviável se houver uma grande quantidade de erros que consumam muito tempo na revisão e correção.

Avaliações de tradutores automáticos permitem que sejam identificadas falhas com a finalidade de aprimorar técnicas e ferramentas de tradução automática ou até mesmo auxiliar usuários na escolha de qual tradutor atende melhor as suas necessidades. Essas avaliações podem ser manuais ou automáticas, uma vez que, as avaliações automáticas possibilitam uma economia maior de tempo e custo. Para a realização de avaliação automática foram desenvolvidas métricas baseadas em tradução humana de referência. Dentre elas, a métrica BLEU é a mais adotada em trabalhos avaliativos.

Este artigo apresenta os resultados obtidos pela avaliação comparativa das ferramentas de tradução online Bing Tradutor e Google Tradutor com o uso de um *corpus* construído para fins de utilização desse trabalho.

2. Tradução Automática

Apesar da existência de pesquisas em tradução automática desde a década de 1950, os sistemas de tradução automática atuais ainda não estão capacitados para substituir tradutores humanos (Martins, 2011)

As expectativas acerca da qualidade das traduções automáticas foram reduzidas ao longo dos anos de pesquisa. Para garantir a qualidade nas traduções geradas automaticamente, é necessário que haja revisão humana. Para Hutchins e Sommers (1992), o maior obstáculo da tradução automática não é computacional, mas, linguístico. A ineficiência nesse processo ocorre porque a tradução depende de conhecimentos semânticos, gramaticais, léxicos, de aspectos extralinguísticos e conhecimentos empíricos. (Sales, 2011). As máquinas ainda não conseguem reproduzir essas características humanas tão peculiares.

A utilização da tradução automática pode se tornar inviável se gerar uma grande quantidade de erros que consumam muito tempo na revisão e correção. A avaliação de

tradutores automáticos pode ser realizada manual ou automaticamente. Porém, avaliações manuais são mais extensas, caras e demandam mais tempo (Papineni et al., 2002). A avaliação automática de tradutores favorece uma forma mais barata e rápida de avaliar a qualidade de traduções do que as avaliações realizadas por julgamentos humanos (Song, Cohn & Specia, 2013).

Para a avaliação automática, as traduções geradas por sistemas são comparadas com traduções humanas de referência. Quanto mais próxima estiver de uma tradução humana, melhor será a qualidade dessa tradução automática. As métricas *Word Error Rate* (WER), *Position Independent Word Error Rate* (PER), *Bilingual Evaluation Understudy* (BLEU), NIST, *General Text Matcher* (GTM) e *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) foram desenvolvidas para aplicar essa concepção.

A métrica BLEU, detalhada na Seção 5, foi adotada para o desenvolvimento do estudo de comparação da precisão dos tradutores automáticos Bing e Google.

3. Tradutores Automáticos Online Avaliados no Estudo

O Bing Tradutor e o Google Tradutor são ferramentas bastante utilizadas pela comunidade web, segundo Tripathi e Sarkhel (2010).

3.1. Bing Tradutor

O Bing Tradutor¹ é um serviço de tradução *online* e gratuito fornecido pela Microsoft®. Inicialmente conhecido como *Microsoft Translator* e, mais tarde, como *Windows Live Translator*, foi desenvolvido para atender necessidades internas da empresa. Em 2007, o produto foi disponibilizado na internet para o público geral. O tradutor da Microsoft® substituiu o Yahoo! Babel Fish, após uma parceria firmada entre as empresas Microsoft® e Yahoo!, em 2012 (Microsoft Translator, 2012).

O Bing Tradutor possibilita a tradução de palavras, frases, textos e páginas da internet, por meio da abordagem de tradução automática estatística. As traduções fornecidas são baseadas nas grandes quantidades de textos armazenados no sistema de tradução da Microsoft®. Atualmente, possui suporte a 45 idiomas.

3.2. Google Tradutor

A empresa Google Inc. oferece traduções instantâneas de forma online e gratuita por meio do Google Tradutor². Permite a tradução de palavras, frases, textos ou páginas da internet.

A princípio, o tradutor do Google era uma versão do tradutor Systran, baseado em regras. Em 2007, a empresa lançou seu próprio sistema de tradução utilizando a abordagem estatística (Korosec, 2011).

O Google Tradutor baseia-se em padrões significativos estatisticamente encontrados em documentos traduzidos por humanos. Estes documentos são textos

¹ <http://www.bing.com/translator/>

² <http://translate.google.com/>

disponíveis na internet por fontes consideradas confiáveis, como textos oficiais de organizações como a Organização das Nações Unidas (ONU), documentos bilíngues de arquivos públicos, livros ou sítios disponíveis na internet. Até o presente momento, o Google Tradutor suporta 80 idiomas

4. Corpus Linguístico

Sardinha (2000) define *corpus* linguístico, ou plural *corpora*, como uma coleção de textos em linguagem natural selecionados e organizados de forma criteriosa para utilização em pesquisas.

Para a avaliação realizada neste trabalho, foi construído um *corpus* de teste paralelo, bilíngue, unidirecional e escrito. O *corpus* é considerado de teste, pois, foi construído com a finalidade de ser utilizado em ferramentas de análise. Bilíngue, por ser composto de textos no par de idiomas inglês-português. Escrito, por ser formado unicamente por textos escritos. E, paralelo, por ser constituído de textos em uma linguagem fonte alinhados a suas respectivas traduções.

Dentro dos critérios definidos para a seleção do *corpus* destaca-se que os textos deveriam ser coletados em endereços eletrônicos confiáveis, possuir o inglês como idioma de origem e sua tradução correspondente em português.

Após a coleta dos textos, foi realizada uma pré-formatação para retirada de conteúdos irrelevantes como cabeçalhos, rodapés, figuras, diagramas etc. Posteriormente, foi feita a tradução de sentença por sentença, do inglês para o português com a utilização dos tradutores Bing e Google. Em seguida, todos os textos foram codificados para UTF-8 e colocados em formato XML (*Extensible Markup Language*).

5. Métrica BLEU

A métrica BLEU (*Bilingual Evaluation Understudy*), é uma das técnicas mais utilizadas para a avaliação de tradutores automáticos (Silva, 2010).

BLEU foi desenvolvida pela *International Business Machines* – IBM, com a finalidade de avaliar sistemas de tradução automática com maior economia, rapidez e independência de linguagens do que avaliações realizadas manualmente. É fundamentada na proximidade entre a tradução automática e a tradução realizada por uma pessoa qualificada em traduções profissionais, chamada de tradução de referência.

A métrica BLEU (Papineni et al., 2002) avalia as traduções automáticas mediante a precisão das sequências de n letras ou palavras, denominadas n -gramas. A precisão de n -gramas indica a proporção de n -gramas compatíveis entre a sentença a ser avaliada, chamada de sentença candidata, e a sentença de referência equivalente. Essa compatibilidade deve ser contabilizada apenas uma vez, recebendo o nome de precisão modificada de n -gramas.

Quando a sentença candidata é maior do que a referência correspondente é atribuída uma penalidade por brevidade (BP), calculada como mostra a Equação 1, pois a sentença candidata deve ser semelhante à referência em tamanho, escolha e ordem de palavras. A penalidade por brevidade é calculada sobre todo o *corpus* e não sobre as sentenças.

$$BP = \begin{cases} 1 & \text{se } c > r \\ e^{(1-\frac{r}{c})} & \text{se } c \leq r \end{cases} \quad (1)$$

Onde, r é a quantidade de palavras do texto de referência e c é a quantidade de palavras do texto candidato.

Finalmente, para calcular a métrica BLEU, calcula-se a média geométrica da precisão modificada, multiplicando-a pelo fator de penalidade por brevidade, conforme apresentado na Equação 2:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N \omega_n \log p_n\right) \quad (2)$$

Em que, BP é o valor de brevidade do texto candidato, P_n é a precisão modificada de n -gramas, N é o limite superior dos n -gramas e ω_n é o peso do n -grama. Considera-se $N=4$ e $\omega_n = 1/N$.

A pontuação BLEU varia de 0 a 1 e quanto mais próximo de 1, mais precisa é considerada a sentença candidata.

6. Avaliação Automática dos Tradutores online Bing Tradutor e Google Tradutor utilizando a métrica BLEU

Por meio da métrica BLEU, foi avaliada a precisão das traduções fornecidas pelos serviços de tradução Bing e Google Tradutor.

Foram selecionados 3 textos dos gêneros jornalístico, técnico e literário: O artigo “Barroso: *Europe is not the cause of the crisis – it’s part of the solution*”, extraído do site do Parlamento Europeu³; o Manual do Usuário da Sony Vaio® para a série VPCF110, retirado da página de suporte da Sony; e um trecho do livro “*Eat, Pray, Love*”, de Elizabeth Gilbert. A quantidade de palavras contidas no *corpus* está representada na Tabela 1.

Tabela 1. Quantidade de palavras contidas no *corpus*

	Documento 1	Documento 2	Documento 3
Fonte	577	213	414
Referência	558	210	417
Bing Tradutor	573	216	423
Google Tradutor	547	215	401

³ <http://www.europarl.europa.eu/portal/pt>

Total	2255	854	1655
-------	------	-----	------

Ao total, o corpus é composto por 4764 palavras em 240 sentenças distribuídas entre os documentos fonte, as traduções de referência e as traduções automáticas.

Para aplicar a métrica BLEU, utilizou-se o MT-Eval, script em Perl desenvolvido pelo National Institute of Standards and Technology – NIST, órgão regulador norte-americano responsável pelo desenvolvimento e fomento de padrões de medidas. O MT-Eval implementa a métrica BLEU e fornece os resultados por documento, por segmento e por sistema mediante do cálculo de n-gramas individuais e cumulativas. As n-gramas individuais referem-se apenas ao cálculo da n-grama sobre o valor assumido por n e n-gramas cumulativas são as médias geométrica de entre todas as n-gramas.

7. Resultados da Avaliação

7.1. Avaliação por Documento

Foram analisados 3 documentos, totalizando 60 sentenças no par de idiomas de inglês-português (en-pt).

Documento 1 – “Barroso: *Europe is not the cause of the crisis – it’s part of the solution*”

O primeiro documento analisado foi um artigo jornalístico retirado do site do Parlamento Europeu, constituído por 577 palavras no texto original, distribuídas em 30 segmentos (sentenças).

As pontuações BLEU obtidas por segmento, calculadas sobre 4-gramas podem ser observadas e comparadas na Figura 1.

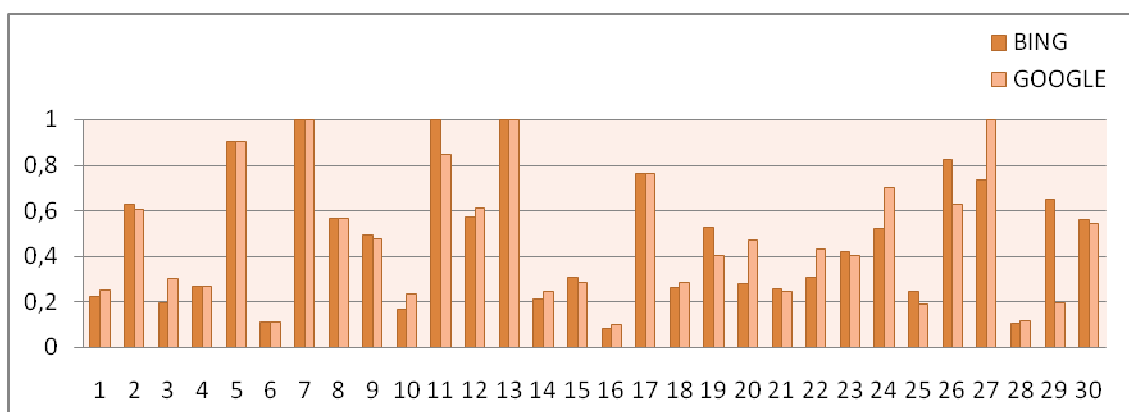


Figura 1. Documento 1 - Comparativo de pontuações BLEU por segmento

Os tradutores Bing e Google obtiveram pontuação igual em 24% dos segmentos. Em 36% o Bing alcançou valores maiores do que o seu concorrente, Google, que se mostrou superior em 40% dos casos.

Apesar da proximidade nos valores obtidos por cada segmento, o Google Tradutor apresentou melhor desempenho do que seu concorrente Bing na avaliação do documento, com a pontuação final de 0.4739. O Bing Tradutor obteve 0.4678 pontos BLEU.

Documento 2 - Manual do Usuário Sony Vaio®, série VPCF110

O documento 2 é um trecho do Manual do Usuário Sony Vaio® para a série VPCF110, retirado do sítio do fabricante, composto por 10 segmentos e 213 palavras no texto de origem.

Utilizando 4-gramas como referência para os cálculos da métrica BLEU, os valores obtidos pelos tradutores avaliados por segmento são comparados na Figura 2.

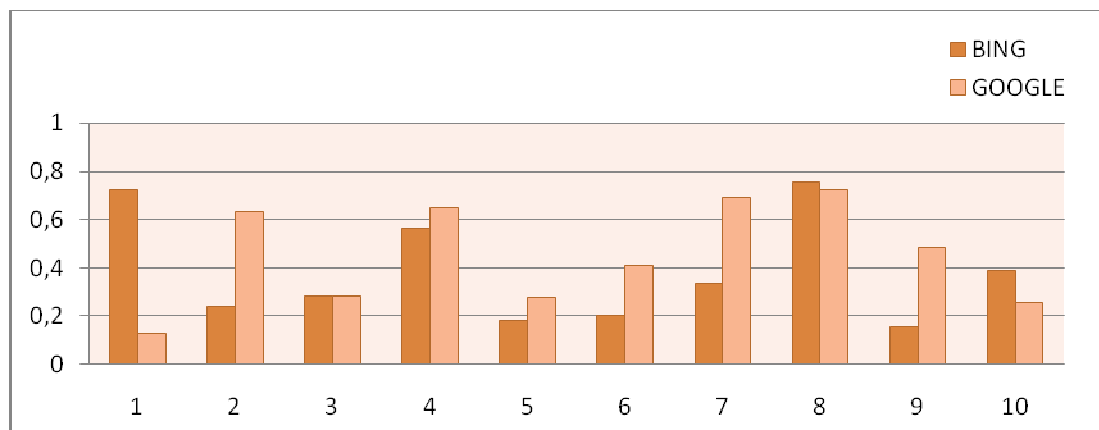


Figura 2. Documento 2 – Comparativo de pontuações BLEU por segmento

O Google apresentou traduções mais precisas em 70% dos segmentos, enquanto o Bing obteve melhor pontuação em 30%.

O melhor desempenho do Google Tradutor na avaliação por segmento ocasionou o melhor desempenho também na avaliação do documento. O tradutor Google obteve a precisão de 48.05%, enquanto o Bing atingiu 34.84% de precisão para o mesmo documento.

Documento 3 – Trecho do livro “Eat, Pray, Love”, de Elizabeth Gilbert.

Documento do gênero literário composto por um trecho retirado do capítulo 3 do livro “Eat, Pray, Love”, de Elizabeth Gilbert. Composto por 20 sentenças, com 414 palavras no texto fonte.

A comparação dos valores obtidos por segmento pelos tradutores Bing e Google pode ser observada na Figura 3.

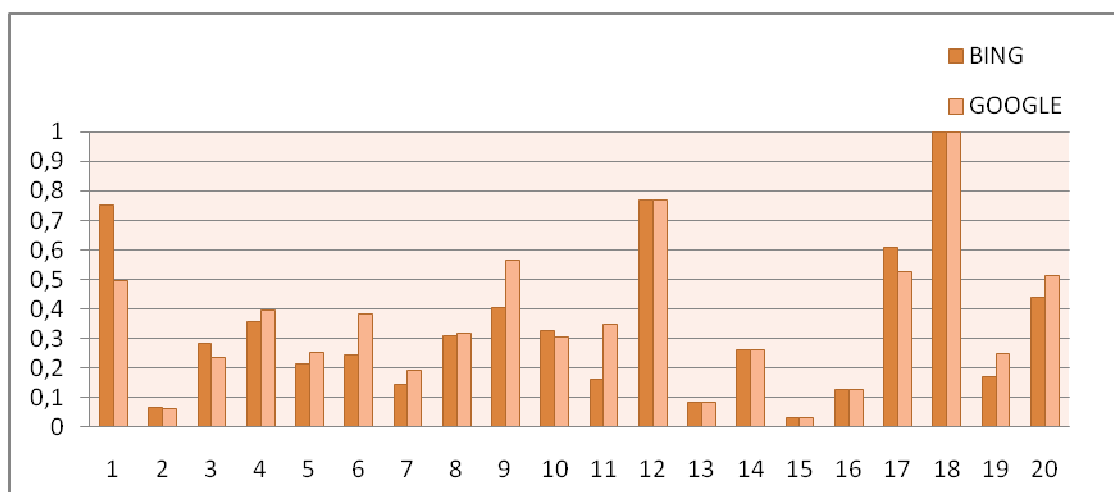


Figura 3. Documento 3 – Comparativo de pontuações BLEU por segmento

O Google foi mais preciso em 45% dos segmentos e o Bing em 30% deles. Em 25% dos segmentos os dois tradutores tiveram o mesmo desempenho.

Na avaliação do documento, o Bing Tradutor alcançou desempenho menor do que o Google Tradutor, com a pontuação BLEU de 0.3394, enquanto seu concorrente atingiu 0.3637.

7.2 Avaliações por Sistema de Tradução

Foram analisados 3 documentos no par de idiomas de inglês-português (en-pt), compostos por 60 sentenças, 1204 palavras nos textos fonte, 1185 palavras nas traduções de referência, 1212 palavras na tradução do Google Tradutor e 1163 na tradução do Bing Tradutor, totalizando 4764 palavras contidas no *corpus* (Tabela 1).

Ao avaliar a precisão por n-gramas individuais (Tabela 2), nota-se que o Google Tradutor apresentou uma pontuação menor do que o Bing Tradutor em apenas em 1-grama individual.

Tabela 2. Precisão das traduções por n-gramas individuais

	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram
Bing	0,7034	0,466	0,3413	0,2574	0,1907	0,1447	0,1078	0,0786	0,0589
Google	0,7013	0,4905	0,37	0,2868	0,2222	0,1718	0,1319	0,1022	0,079

As pontuações BLEU atingidas pelos dois tradutores por n-gramas cumulativas são mostradas na Tabela 3.

Tabela 3. Precisão das traduções por n-gramas cumulativas

	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram
Bing	0,6902	0,5618	0,4728	0,4042	0,3465	0,2986	0,2575	0,2214	0,1907
Google	0,7013	0,5865	0,503	0,4371	0,3818	0,3342	0,2926	0,2566	0,2251

Para n-gramas cumulativas, o Google Tradutor destacou-se para todos os valores assumidos por n.

A avaliação final dos sistemas de tradução, baseada no *corpus* definido para esse trabalho e anteriormente avaliado, infere que o Google Tradutor, com 43.71% de precisão, apresentou melhores traduções do que o Bing Tradutor, com 40.42% (Figura 4).

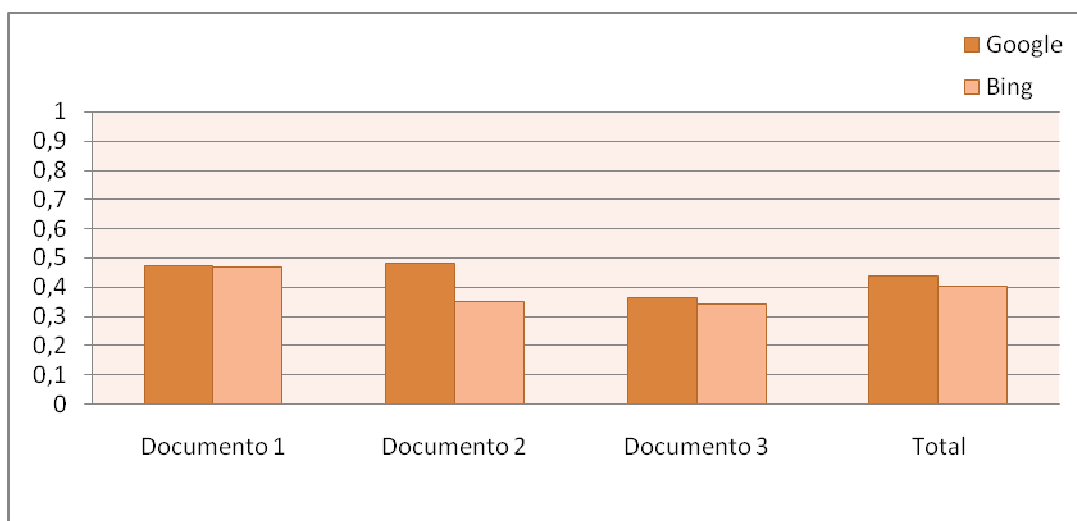


Figura 4. Comparativo da pontuação final BLEU por sistema de tradução

A diferença numérica entre a pontuação dos dois tradutores, de acordo com a métrica BLEU foi de apenas 0,0329 pontos.

8. Conclusão

A avaliação de tradutores automáticos é uma parte importante da Tradução Automática. Por meio de avaliações é possível observar a precisão alcançada e identificar os erros mais comuns cometidos pelos tradutores para buscar o aprimoramento constante dos sistemas.

Nesse artigo, foram avaliados e comparados, de forma quantitativa, os tradutores da empresa Microsoft®, Bing Tradutor, e o tradutor da empresa Google Inc., Google Tradutor. A métrica de avaliação BLEU foi aplicada nos resultados gerados por esses tradutores com a utilização do script MT-Eval e de um *corpus* construído para utilização nesse estudo.

Embora o tradutor da Google tenha apresentado uma precisão melhor em grande parte dos segmentos avaliados, pode-se notar que a diferença final na pontuação BLEU foi pequena, sendo 0,0329 pontos. Isso reflete o desempenho similar de ambos os tradutores na análise feita por segmento e por documento.

Assim como a tradução automática ainda não é um processo independente da intervenção humana, a sua avaliação também não é. Ressalta-se que a avaliação pela métrica BLEU calcula a pontuação por intermédio da comparação entre traduções humanas de referências e as traduções automáticas, ou seja, o resultado depende da similitude entre essas traduções.

Com o entendimento e a aplicação da métrica na avaliação das traduções desse trabalho, foi verificado que o estilo de tradução, a escolha por palavras sinônimas, o conhecimento do contexto e a ordem das palavras diferentes nas sentenças de tradutores humanos e automáticos influenciaram a pontuação BLEU. Porém, isso não significa que a tradução automática seja de má qualidade. Esse fato se justifica porque uma única sentença pode possuir várias traduções possíveis, diferentes entre si e com o mesmo significado. Os fatores citados não podem contribuir para julgar uma tradução como inadequada. A avaliação de tradução automática é uma tarefa abstrata. Por ser uma métrica, a BLEU é apenas uma forma quantitativa de avaliar a precisão das traduções e é insuficiente realizar avaliação de traduções com a utilização de métricas, uma vez que lingüística não é uma ciência exata e as linguagens passam por constantes modificações.

Referências Bibliográficas

HUTCHINS, J.; SOMERS, H. L. (1992). *An introduction to machine translation*. Academic Press. Em: <http://www.hutchinsweb.me.uk/IntroMT-TOC.htm>. Acesso em 26 de maio de 2014.

KOROSEC, M. K. (2011). *The Internet, Google Translate and Google Translator Toolkit Nuisance or Necessity in Translator Training?* Tralogy. Em: <http://odel.irevues.inist.fr/tralogy/index.php?id=113>. Acesso em 25 de junho de 2014.

MARTINS, R. (2011). O pecado original da tradução automática. *Alfa*, Universidade Estadual de São Paulo. Em: <http://seer.fclar.unesp.br/alfa/article/viewFile/4178/3776>. Acesso em 10 de agosto de 2014.

Microsoft Translator, Microsoft. (2012). Welcome Yahoo! Babel Fish users. Em: <http://blogs.msdn.com/b/translation/archive/2012/05/30/welcoming-yahoo-babel-fish-users.aspx>. Acesso em 25 de julho de 2014.

PAPINENI, K.; S. ROUKOS, T. W.; Zhu, W. (2002). BLEU: a method for automatic evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, p. 311–318.

SALES, S. G. (2011). Tradução Automática: Os processos da tradução mediada por computador. *Saberes em perspectiva*, Jequié: Universidade Estadual do Sudoeste da Bahia, v.1 n.1 p. 19-37.

SARDINHA, T. B. (2000). Linguística de Corpus: Histórico E Problemática. *Revista D.E.L.T.A*, São Paulo, v. 16 , n. 2. , p. 323-367. Em: <http://ref.scielo.org/5mc5wv> Acesso em 19 de novembro de 2014.

SILVA, F. (2010). *Análise Comparativa dos Resultados de Mecanismos de Tradução Automática Baseados em Regras e Estatística*. Dissertação (Mestrado em Estudos da Tradução) – Universidade Federal de Santa Catarina, Florianópolis.

SONG, X.; COHN T.; SPECIAL. (2013). BLEU deconstructed: Designing a Better MT Evaluation Metric. In: Proceedings of the International conference on Intelligent text processing and computational linguistics, 14., 2013, Samos: *International Journal of Computational Linguistics and Applications*, p. 29-44.

TRIPATHI, S.; SARKHEL, J. K. (2010). Approaches to machine translation. *Annals of library and information studies*, v. 57, p. 388-393. Em: [http://nopr.niscair.res.in/bitstream/123456789/11057/1/ALIS%2057\(4\)%20388-393.pdf](http://nopr.niscair.res.in/bitstream/123456789/11057/1/ALIS%2057(4)%20388-393.pdf) Acesso em 10 de junho de 2014.