e-scrita ISSN 2177-6288

V. 5 – 2014.3 –F. MELO, H. MATOS, E. DIAS

APLICAÇÃO DA MÉTRICA BLEU PARA AVALIAÇÃO COMPARATIVA DOS TRADUTORES AUTOMÁTICOS BING TRADUTOR E GOOGLE TRADUTOR

Francisco Ramos de Melo[1]

Hellen Carmo de Oliveira Matos[2]

Emillie Rebecca Bastos Dias[3]

RESUMO: Este artigo descreve como a métrica BLEU auxilia na avaliação de traduções automáticas e apresenta os resultados obtidos em uma avaliação 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. Para a aplicação da métrica BLEU neste trabalho, foram utilizados textos do *corpus* que foi construído para essa finalidade, e que, após a tradução, serão submetidos à avaliação com a utilização da métrica implementada pelo *script* MT-Eval. Ao final, as pontuações resultantes da avaliação entre os tradutores selecionados são apresentadas de forma comparativa.

Palavras-chave: avaliação de tradução automática; BLEU; MT-Eval.

APLICATION OF THE BLEU SCORES TO COMPARATIVE EVALUATION OF AUTOMATIC TRANSLATORS BING AND GOOGLE TRANSLATOR

ABSTRACT: This paper describes how the BLEU scores are used by Machine Translation Evaluation, assesses the accuracy of Bing Translator and Google Translate and gives a

[1] Doutor (2012) e Pós Doutor (2013) em Ciências pela Universidade Federal de Uberlândia e Mestre em Engenharia Elétrica e de Computação pela Universidade Federal de Goiás na linha de Inteligência Artificial. GO, Brasil. francisco.melo@ueg.br

[2] Professora com Dedicação Exclusiva da Universidade Estadual de Goiás. Mestre em Ciência da Computação pela Universidade Federal de Goiás. GO, Brasil. hellen.sistemas@gmail.com

[3] Graduanda em Sistemas de Informação pela Universidade Estadual de Goiás. Servidora pública municipal da cidade de Anápolis, Goiás. Ex-bolsista (CNPq) de graduação sanduíche na Universidade Católica de Ávila, Espanha. emillie.info@gmail.com

comparison between the outputs. The BLEU score is widely used method to evaluate the accuracy of machine translations using human reference translations. The texts used in this work comprise the corpus that was constructed for this purpose, and which, after translation, will be evaluated by using the BLEU metric implemented by the MT-Eval script. Finally, the scores from the evaluation of selected translators will be provided for comparative purposes.

Key words: Machine Translation Evaluation, BLEU, MT-Eval.

1. Introdução

A tradução automática é o processo de utilização da computação para converter uma mensagem de uma linguagem natural a outra, mantendo a equivalência com o conteúdo original. Os primeiros projetos nessa área foram desenvolvidos com motivações militares durante a Guerra Fria e, desde então, a tradução automática tornou-se um meio facilitador de relações socioeconômicas, 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 estudos 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 confiabilidade nesse tipo de tradução ainda depende da revisão humana para garantir o quão utilizável o texto traduzido pode ser. A aplicação da tradução automática pode se tornar inviável se houver uma grande quantidade de erros que consumam muito tempo na revisão e correção.

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.

As 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 descreve como a métrica BLEU é utilizada para avaliações de traduções automáticas, e apresenta os resultados obtidos pela avaliação comparativa das ferramentas de tradução *online* Bing Tradutor e Google Tradutor, através 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, p. 287)

As expectativas acerca da qualidade das traduções automáticas foram reduzidas ao longo dos anos de pesquisa. Não é possível confiar totalmente nas saídas automáticas, sem que haja uma revisão humana que garanta a qualidade da tradução gerada (HUTCHINS; SOMMERS, 1992). A ineficiência nesse processo ocorre porque a tradução depende de conhecimentos semânticos, gramaticais, léxicos, de aspectos extralinguísticos e conhecimentos empíricos (SALES, 2011, pp. 19-37). 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.

Para a realização de avaliação automática foram desenvolvidas métricas baseadas em tradução humana de referência. A métrica BLEU (*Bilingual Evaluation Understudy*), é uma das técnicas mais utilizadas para a avaliação de tradutores automáticos (BURCH, OSBORN, KOEHN, 2006; JONES, SHEN E HERZOG, 2009; SILVA, 2010).

3. BLEU

BLEU é uma métrica desenvolvida pela *International Business Machines* ï 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 avalia as traduções automáticas através da precisão das sequências de *n* letras ou palavras, denominadas *n-gramas*. A precisão de *n-gramas* indica a quantidade de *n-gramas* compatíveis entre a sentença a ser avaliada, chamada de sentença candidata, e a sentença de referência equivalente, dividindo este número pelo total de palavras da sentença candidata. 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, calculada como mostra (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.

$$BP = \begin{cases} 1 & \text{se c} > \text{r} \\ \exp\left(1 - \frac{r}{c}\right) & \text{se c} = \text{r} \end{cases} \qquad (1)$$

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 mostra a fórmula (2):

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

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 w_n é o peso do *n-grama*. Considera-se N=4 e $w_n = 1/N$.

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

4. Avaliação Automática dos Tradutores *Online* Bing Tradutor e Google Tradutor utilizando a métrica BLEU

Através da métrica BLEU, foi avaliada a precisão das traduções fornecidas pelos serviços de tradução amplamente utilizados por usuários da internet, Bing Tradutor e Google Tradutor (TRIPATHI E SARKHEL, 2010). Atualmente, o Bing suporta 45 idiomas, enquanto o seu concorrente, Google, possui suporte a 80 idiomas.

Para a avaliação, 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.

Os documentos para composição do *corpus* foram selecionados em endereços eletrônicos considerados confiáveis. Foram selecionados 3 textos dos gêneros jornalístico, técnico e literário: O artigo ñBarroso: *Europe is not the cause of the crisis ï it ôs part of the solution*ò , extraído do *site* do Parlamento Europeu; o Manual do Usuário da Sony Vaio® para a série VPCF110, retirado da página de suporte da Sony; e um trecho do livro ñ*Eat, Pray, Love*ò, 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
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.

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, através dos tradutores Bing e Google. Em seguida, todos os textos foram codificados para UTF-8 e colocados em formato XML (*Extensible Markup Language*).

Para aplicar a métrica BLEU, utilizou-se o MT-Eval, *script* em Perl desenvolvido pelo *National Institute of Standards and Technology* ï 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 através do cálculo de *n-gramas* individuais e cumulativas. Os *n-gramas* individuais referem-se apenas ao cálculo do *n-grama* sobre o valor assumido por *n* e *n-gramas* cumulativos são as médias geométrica de entre todos os *n-gramas*.

5. Resultados da Avaliação

5.1 Avaliação por Documento

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

Documento 1 ï ñBarroso: *Europe is not the cause of the crisis ï itôs part of the solutionò*

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.

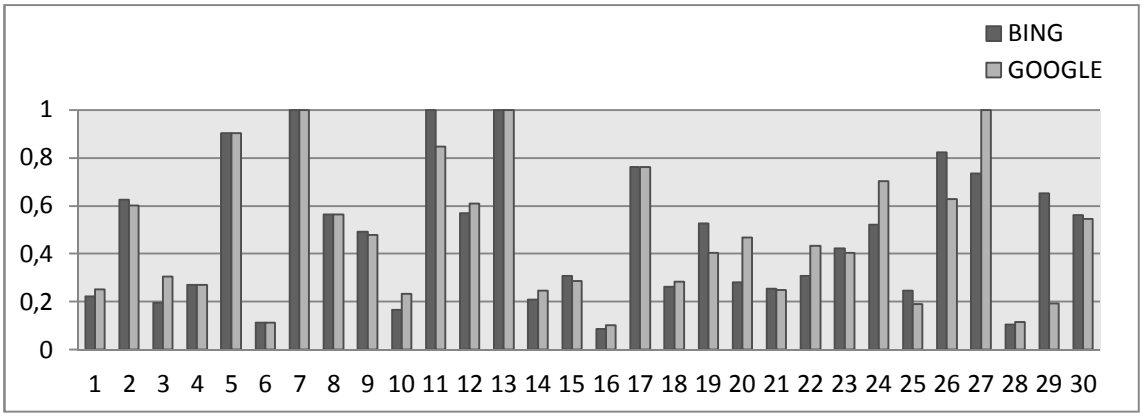

Figura 1. Documento 1 - Comparativo de Pontuações BLEU por segmento

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 Usúario 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.

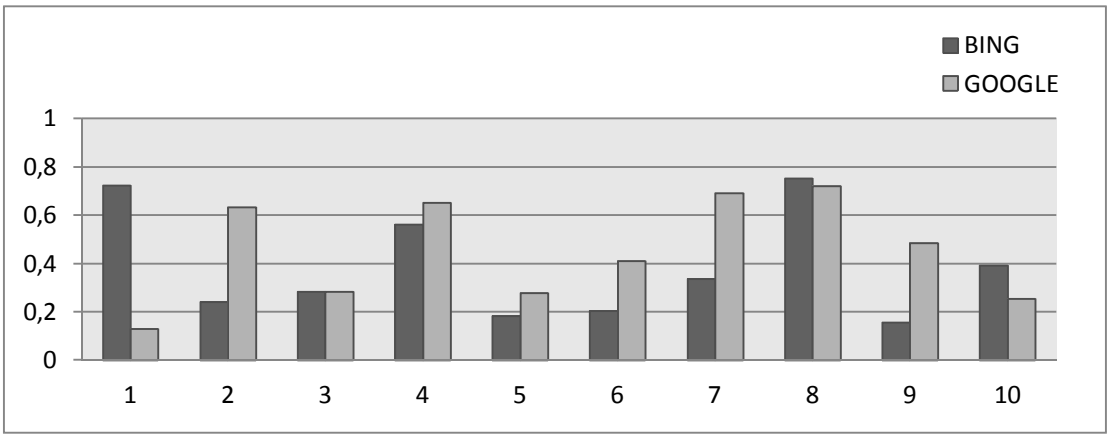

Figura 2. Documento 2 ï Comparativo de Pontuações BLEU por segmento

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.

Documento3 ï Trecho do livro ñ*Eat, Pray, Love*ò, de Elizabeth Gilbert.

Documento do gênero literário composto por um trecho retirado do capítulo 3 do livro ñ*Eat, Pray, Love*ò, 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.

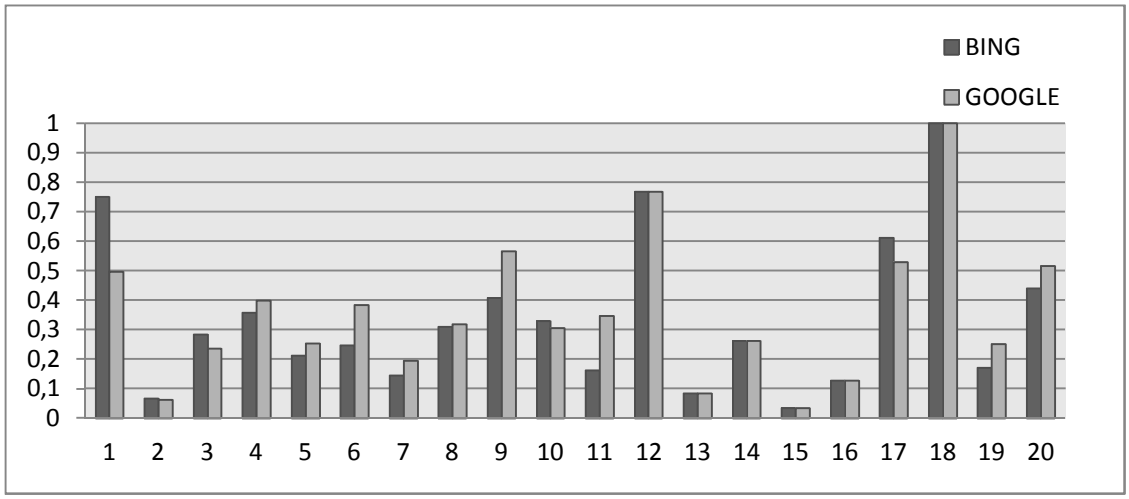

Figura 3. Documento 3 ï Comparativo de Pontuações BLEU por segmento

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.

5.2 Avaliação 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 *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* cumulativos são mostradas na tabela 3.

Tabela 3. Precisão das traduções por *n-gramas* cumulativos

	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* cumulativos, 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).

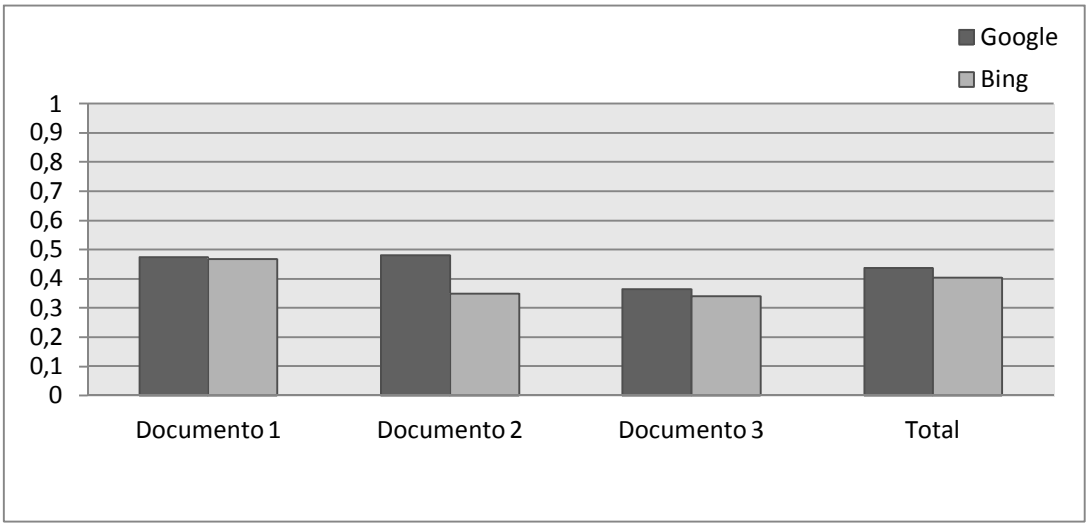

Figura 4. Comparativo da pontuação final BLEU por sistema de tradução

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

6. Conclusão da Avaliação

A avaliação de tradutores automáticos é uma parte importante da Tradução Automática. Através dela 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 através da utilização do *script* MT-Eval e de um *corpus* construído para utilização nesse estudo.

Com a análise dos resultados obtidos, pode-se notar que a diferença final na pontuação BLEU dos tradutores avaliados foi pequena, sendo 0.0329 pontos. Isso reflete o desempenho similar de ambos os tradutores na análise feita por segmento e por documento. Os erros cometidos pelos tradutores também são semelhantes: Tempos verbais, conjugações incorretas, palavras ausentes ou extras, artigos/pronomes ausentes, ocultos ou extras, erros semânticos, dentre outros. Embora seja pequena, a diferença no resultado final, tanto nas pontuações por documentos quanto na pontuação total favorece o Google Tradutor. E, de fato, por meio dessa avaliação comparativa, o tradutor da Google apresentou precisão melhor durante os testes com o *corpus* desse trabalho.

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 através 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 prejudicaram 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 apenas através de métricas, uma vez que linguística não é uma ciência exata e as linguagens passam por constantes modificações.

REFERÊNCIAS BIBLIOGRÁFICAS

BURCH, C. C.; OSBORNE, M. KOEHN, P. Re-evaluating the Role of BLEU in Machine Translation Research. *In Proceedings of EACL*, Conference of the European Chapter of the Association for Computational Linguistics. 2006.

HUTCHINS, J.; SOMERS, H. L. *An introduction to machine translation*. Manchester: London: Academic Press, 1992.

JONES, D.; SHEN, W.; HERZOG M. Machine Translation for Government Applications. Lincoln Laboratory Journal, n. 1, v. 18, 2009.

MARTINS, R. O pecado original da tradução automática. *Alfa*, Araraquara: Universidade Estadual de São Paulo, 2011.

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

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

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

TRIPATHI, S.; SARKHEL, J. K. Approaches to machine translation. *Annals of library and information studies*, v. 57, pp. 388-393, 2010.

Recebido em 21/11/2014.

Aceito em 13/12/2014.