Correlation between Automatic Evaluation Metric Scores, Post-Editing Speed, and Some Other Factors

Midori Tatsumi

School of Applied Language and Intercultural Studies
Dublin City University
Glasnevin, Dublin 9, Ireland
Midori.tatsumi2@mail.dcu.ie

Abstract

This paper summarises the results of a pilot project conducted to investigate the correlation between automatic evaluation metric scores and post-editing speed on a segment by segment basis. Firstly, the results from the comparison of various automatic metrics and post-editing speed will be reported. Secondly, further analysis is carried out by taking into consideration other relevant variables, such as text length and structures, and by means of multiple regression. It has been found that different automatic metrics achieve different levels and types of correlation with post-editing speed. We suggest that some of the source text characteristics and machine translation errors may be able to account for the gap between the automatic metric scores and post-editing speed, and may also help with understanding human post-editing process.

1 Introduction

While machine translation (MT) is becoming prevalent in large scale translation production environments in the hope of reducing costs and improving efficiency, MT output quality is not yet considered good enough to achieve effective use of MT, and thus human post-editing (PE) is still considered as a demanded process (TAUS, 2009). As the time and cost of human PE can sometimes be prohibitively high, the effectiveness and reduction of the human PE process should remain of interest to commercial users where fully automated MT is not the norm. Some research has been conducted to measure PE effort. Krings (2001) employed Think Aloud Protocol, where post-editors were asked to verbalise all thoughts they had during the PE task in order to investigate post-editors' cognitive effort. O'Brien (2006a, 2006b) introduced Translog and an Eye-Tracker to record the keyboard operation and eye movement to examine post-editors' cognitive and technical effort. These extensive qualitative analyses have helped with understanding the PE task more fully.

However, a quicker and easier method for overall analysis may be in demand for commercial users, who constantly deal with projects that require translation of hundreds of thousands of words. One of the possible methods is using automatic evaluation metrics. As these metrics compare two translations and calculate the distance, the textual differences made during the PE process can be easily measured. However, a question remains as to how precisely the textual difference represents the actual PE effort; some errors can be almost instantly correctable while others may need longer consideration. In fact, Krings (ibid) has suggested that the textual difference may not always reflect the actual amount of PE effort.

One of the main purposes of this paper is to examine how well automatic metric scores correlate the amount of PE effort. As a method to capture the amount of PE effort, we measure PE time. Time measurement could be a useful method especially in the commercial context. A time keeping function can be relatively easily embedded in professional post-editors' standard work environment, thus enabling the capture of real-life data rather than conducting the experiment in a lab. Time is a simple numerical measure and the data from large samples can be processed and analysed with relative ease. Time affects the production schedule and cost, and is therefore relevant to the industry. The examination of the correlation between automatic metric scores and PE time may give us some insight into understanding the PE effort in terms of the relationship between product (textual difference) and process (effort).

While the correlation between automatic evaluation metric scores and human evaluation scores have extensively been researched and reported (Papineni et al, 2002) (Turian et al, 2003) (Snover et al, 2006, Callison-Burch et al, 2008), the correlation between automatic metric scores and human PE time has received little attention so far. Guerberof (2008) examined the time data in order to assess the realistic price setting for the post-editing of MT output in relation to editing of translation memory fuzzy matches, but automatic metrics were not taken into consideration.

Another purpose of this paper is further analyse the relationship between automatic metric scores and PE time by taking into consideration additional variables that may increase or decrease the amount of PE effort in order to gain an insight into the nature of PE tasks. We employ multiple regression as a main analysis method. Multiple regression makes it possible to conduct extensive quantitative analyses by including a number of variables at a time, which may be helpful in breaking down the human effort and tasks in the PE process.

The reminder of this paper is organised as follows. In section 2, the experimental setting will be explained. In section 3, the results of a comparison between four automatic metrics in terms of correlation with PE time data will be presented. In section 4, further explanatory variables will be taken into consideration in multiple regression models, and the results will be discussed, and section 5 concludes the paper.

This project has been conducted on English to Japanese translation, but the logic and the methodology may be applicable to other language pairs.

2 Experimental Settings

2.1 Test corpus

This project has been conducted in collaboration with Symantec Corporation, and the test corpus was compiled from the documentation of one of their recent computer security products released earlier in 2009. The source text was written to conform to Symantec's controlled language rules. The test corpus consisted of 4,784 words in 475 segments.

2.2 MT and PE software

The test corpus was translated by Systran version 5, customised by Symantec's user dictionaries. The text was processed using Symantec's pre- and post- processing scripts, which perform mainly global search and replace to make the source text more amenable for MT and to repair the target text as much as possible before the human PE process.

The PE was done using SDL Trados Translator's Workbench and TagEditor, one of the industry standard translation memory (TM) and PE tools. The time was recorded by means of the standard feature of Trados combined with a Windows macro devised for this project to achieve more thorough time measurements.

2.3 Post-Editors

Three professional Japanese native-speaking translators were employed for the task, and each of them post-edited the entire test corpus. Brief guidelines were provided mainly to emphasise the quality requirements for post-edited text; it has to convey correctly the meaning of the source text, and conform to the Japanese grammar, but does not have to be stylistically sophisticated. After the PE, the word count of each source segment was divided by the time taken to edit the corresponding segment to obtain a word per minute speed.

Two of the translators had more than 10 years of experience both in translation in the relevant subject and with the TM tools, and one of them also had some experience in PE (less than one year). The other translator only had experience in translation in unrelated subjects and 1-3 years of experience with the TM tools.

2.4 Automatic Metrics

Four automatic metrics, namely, GTM (General Text Matcher) (Turian et al, 2003, Melamed et al, 2003), TER (Translation Edit Rate) (Snover et al, 2006), BLEU (BiLingual Evaluation Understudy) (Papineni et al, 2002), and NIST (National Institute of Standards and Technology) (Doddington, 2002) were used to obtain the evaluation scores between MT output and the post-edited final text. The main criteria for choosing the metrics were: 1) applicable for both wide variety of European languages and Japanese, and 2) frequently used in relevant research and literature. BLEU and NIST are also employed in this project despite the fact that they are not designed for sentence level evaluation, since comparison with PE speed is a rather new approach and thus it may be worth testing. As the Japanese writing system does not insert spaces to mark the boundary of words, the text was tokenised by means of ChaSen.[1]

3 Correlation

Table 1 shows the raw speed data for each post-editor. The difference in speed may be mainly due to the difference of experience in translation and tools.

Post-Editor	Mean	Std.Dev.	Min	Max
A	18.08	15.83	0.74	100
B	33.43	21.01	1.43	150
C	36.37	22.10	2.54	150

Table 1. Summary statistics for PE speed (word/min)

Table 2 shows the results for automatic metric scores calculated on a segment by segment basis, from all three post-editors, thus containing 1,425 observations (475 segments edited by three post-editors).

	Mean	Std.Dev.	Min	Max
GTM	0.75	0.21	0	1
TER	28.31	33.06	0	300
BLEU	0.48	0.37	0	1
NIST	8.25	3.04	0	19.52

Table 2. Summary statistics for automatic metric scores (N: 1425)

Figure 1 shows a set of scatter plots that depict the relationship between PE speed (y-axis) and automatic metric scores (x-axis), along with a Pearson correlation coefficient in parentheses above each plot. Since the raw PE speed data have a positively skewed distribution and have a slightly exponential relationship with automatic metric scores, they are transformed to logarithm numbers to ensure a normal distribution and a linear relationship with automatic metric scores. Also, in order to find out the general trend common to all post-editors, the logarithmic PE speed data from each post-editor were converted to Z-scores for calculation of the correlation coefficients. Each data point represents each segment. The graphs have been drawn based on all observations from three post-editors, but the shape of the distribution was similar for each post-editor.

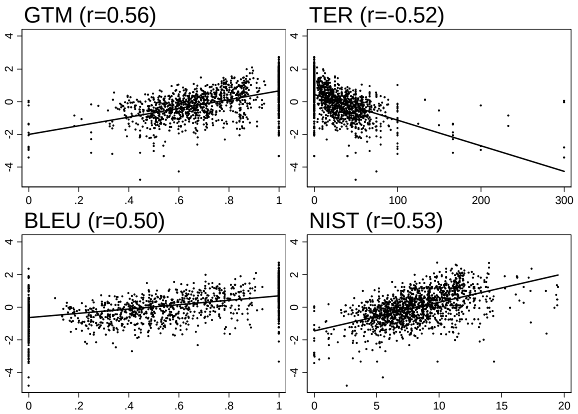

Figure 1. Correlation between automatic metric scores and Z-scores of logarithmic PE speed

The shapes of the distribution are quite characteristic to each metric, with TER having a slope in the opposite direction as it gives a zero score for a perfect match between MT and PE output and increases the score as the distance between the two becomes larger.

The groups of dots on the right edges of GTM and BLEU and the left edge of TER represents the 'perfect match' between MT and PE. As can be seen, the perfect match segments, which have required no PE, show broad range of 'PE speed.' This means that some sentences required the post-editors to spend more time than other sentences to reach the decision that no editing is necessary. It may be interesting to investigate what conditions cause such difference, but further analysis is withheld here as that is out of scope of this paper.

The main reason that BLEU also has the left edge group is that its standard 4-gram option gives a zero score to all segments that are shorter than four words, even if they are a perfect match. While GTM and BLEU accommodates all possible distances within the range between 0 and 1, TER does not have an upper limit (the distance between MT and PE depends partly on the difference in lengths of the two texts), and NIST does not have a limit for a perfect match. Therefore, TER and NIST tend to produce extreme values that can be influential to statistical analysis. For example, the

[1] Copyright by Nara Institute of Science and Technology. Accessible from: http://sourceforge.jp/projects/chasen-legacy/

Pearson correlation coefficient for TER changes from -0.52 to -0.55 if data points with the scores higher than 100 are discarded.

Among all four metrics, GTM has obtained the strongest correlation with PE speed, but the distribution of the data points is still broad. In the next section, we focus on GTM scores and consider some of the explanatory variables that may account for the variance.

4 Possible Explanatory Variables

For the purpose of this paper, a small set of possible explanatory variables will be discussed, which consists of two source text characteristics and one MT error type, namely, source segment length, source segment structure, and dependency error. After a brief explanation of each issue, a multiple regression analysis will be performed incorporating all these variables.

4.1 Source segment length

The segment length often becomes an issue for writing the source text for MT. Extremely short sentences often lack contexts and thus may be semantically ambiguous for both humans and MT. On the other hand, long sentences can entail both grammatical and semantic complexity, which is also problematic for both humans and MT.

Figure 2 shows the relationship between PE speed and source segment length measured by word count. As seen from the distribution and the fitted line, the effect of segment length on PE speed seems to be non-linear.

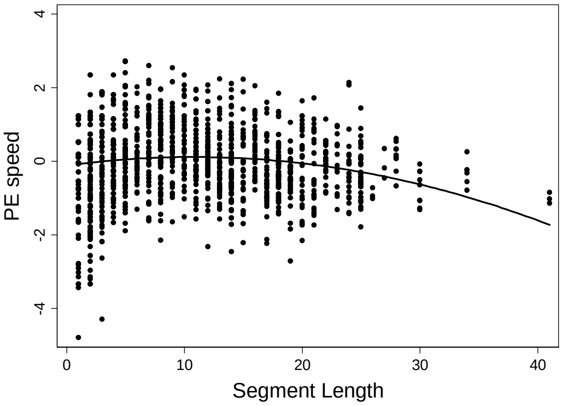

Figure 2. Effect of segment length on PE speed

In order to examine further this relationship while also taking into account other variables, a quadratic specification will be included in the multiple regression analysis.

4.2 Source segment structure

It is understandable that the sentences with more complex structures are more difficult to understand, translate, and edit for humans. Leech (2006) suggests three major categories of English sentence structures: a *simple sentence* contains only one clause, a *compound sentence* contains two or more clauses linked by coordination, such as ‘and’ and ‘but’, while a *complex sentence* contains one or more subordinate clauses. Examples of each category taken from the test corpus are shown below.

Simple sentence:

- An email has more than four attachments.
- To delete items from a vault other than your private vault, you need appropriate access permissions.

Compound sentence:

- The shortcut is a direct link to the archived item, and it has the following icon.

Complex sentence:

- Select the items that XXX is processing.
- Put the item in the Restored Items folder in the mailbox that is specified in the Settings dialog box.

In this paper, we add one more category *incomplete sentence* (words and phrases) to accommodate the characteristics of technical documentation, which tends to contain a number of segments that do not fall into any of the aforementioned sentence categories. In fact, 200 segments out of 475 analysed in this paper are incomplete sentences.

Incomplete sentence:

- File size
- For a file system vault:
- If there is more than one page of search results:

As seen in the examples, some incomplete sentences are easily understood, while others suffer from contextual ambiguity. This is suspected to result in variations in difficulty for both MT and humans.

	Number	Mean	Std. Dev.	Min	Max
Incomplete	200	3.72	2.97	1	18
Simple	143	11.14	5.18	2	25
Compound	4	18.25	6.70	11	27
Complex	128	18.54	5.74	7	41
All	475	10.07	7.63	1	41

Table 3. Segment length (word count) by segment structures

Table 3 shows that the mean segment length increases from incomplete to simple to compound to complex sentences. However, as seen from the above examples, it is not always the case that the longer sentences have more grammatical complexity than shorter sentences, and vice versa. Therefore, it may be worthwhile to take into account such characteristics in addition to the segment length. However, compound sentences will be omitted from the analysis since there are only four observations in the entire corpus, although Symantec's controlled language rules do not explicitly restrict the use of any sentence structures.

Figure 3 shows the GTM vs. PE speed distribution broken down by sentence structures, which illustrates a tighter relationship between PE speed and GTM scores in simple sentences when compared with incomplete and complex sentences. To distinguish the difference, multiple regression will be performed on separate samples of each sentence structure.

4.3 Dependency errors

For the purpose of this paper, 'dependency error' simply means the mistranslation of semantic relationships between words and phrases, including subject-verb confusion, modifier-modified confusion, and so on. Some examples include (gloss is shown in brackets):

EN: *XXX finds only archived emails.*
JA-MT: XXX 発見は電子メールのみをアーカイブしました。
[XXX Finds has archived only emails. ("XXX Finds" is treated as a subject.)]
JA-PE: XXX はアーカイブ済み電子メールのみを検索します。
[XXX searches for only the emails that have been archived.]

EN: *Downloading items to your vault cache*
JA-MT: ボルトキャッシュへのダウンロードのアイテム
[Items for downloading to your vault cache]
JA-PE: ボールトキャッシュへのアイテムのダウンロード
[Downloading items to your vault cache]

It is suspected that correcting dependency errors cause more cognitive effort when compared to word level changes, since sentences that include dependency errors often make sense on the surface but the meaning does not match the source text, thus requiring more PE time to repair the translation. The required textual corrections, however, are not necessarily extensive; it could be a change of one preposition or the position of a phrase, which means such cognitive effort may not be reflected properly in automatic metric scores. For simplicity, a binary independent variable will be considered in the multiple regression analysis; 0

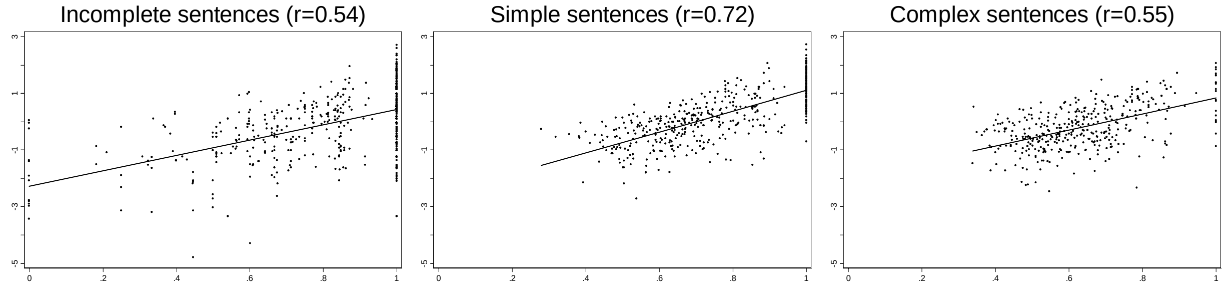

Figure 3. Correlation between GTM scores (x-axis) and PE speed (y-axis) by sentence structures

for segments with no dependency error, and 1 for segments with one or more dependency errors.

4.4 Quantitative analysis with multiple regression

The regression results are reported in Table 4. The effect of each variable on increasing PE speed is shown as unstandardised coefficients as a result of one unit increase of each variable. Statistical significance is shown by asterisks; three asterisks mean that the p-value is less than 0.01. Model I shows the results from the entire sample, while II, III, and IV show the separate results by sentence structures. For the dependent variable, instead of the Z-score of logarithmic PE speed, unstandardised logarithmic PE speed is employed, and additional binary variables for Post-Editor A and B are introduced to control for the difference in overall speed between post-editors.

The values for GTM Scores show how much GTM scores, holding other conditions fixed, relate to PE speed. The PE speed for GTM=1 segments (perfect match), is approximately 5 times, 14 times, and 5.5 times faster compared to GTM=0 segments for incomplete, simple, and complex sentences respectively[2]. If we break down the GTM score into a smaller unit, an increase of 0.1 in the GTM score increases the PE speed by 19.9%, 31.3%, and 20.5% for incomplete, simple, and complex sentences respectively[3]. In either unit, it can be seen that GTM scores have stronger relationship with PE speed for simple sentences compared to incomplete and complex sentences.

The coefficient for Sentence Length for incomplete sentences is positive (0.225) and its squared term is negative (-0.011). This is the evidence of the quadratic, inverted U-curve effect of sentence length on PE speed. According to the estimate shown in the table, the threshold is 10/11 [4]; for segments shorter than 11 words, holding other conditions fixed, the PE speed becomes faster as the segment length approaches to 10 words, and for segments longer than 10 words, the PE speed becomes slower as the segment length becomes longer. A similar effect can be observed for simple sentences, but the evidence is weaker (0.071 and -0.002). The estimate shows that the threshold for simple sentences is 15/16[5]. As for complex sentences, the statistical significance for this variable was not obtained. One of the possible explanations for the inverted U-curve effect may be that the longer segment helps disambiguation up to a certain point, but also increases semantic and/or grammatical complexity after that point onward. To find out why incomplete sentences are affected more strongly compared to simple sentences, a detailed qualitative investigation may be required.

Dependency also seems to have different effects on different sentence structures. According to the estimate, presence of one or more dependency error(s) slows down the PE speed by 32.6% and 21.7% in incomplete and complex sentences respectively, compared with only 2.2% in simple sentences [6]. The result for simple sentences is not statistically significant either. In the case of complex sentences, it is intuitively understandable that dependency errors in sentences whose structure is complex may make PE a much more effort-intensive task than sentences with a simpler structure. In the case of incomplete sentences, however, the inherent contextual ambiguity may be one of the causes for difficulties. In any case, considering the statistical and substantive significance, the dependency issue deserves more extensive analysis both qualitatively and quantitatively, which will be one of the key interests of our future research.

The variables for Post-Editor A and Post-Editor B show the overall speed differences in comparison to Post-Editor C. Since these variables have been introduced only for the purpose of cancelling out the inter-subject differences, further analysis is withheld in this paper.

[2] 100*(exp(1.811)-1)=511, 100*(exp(2.720)-1)=1418, 100*(exp(1.863)-1)=544

[3] 100*(exp(.1811)-1)=19.9, 100*(exp(.2720)-1)=31.3, 100*(exp(.1863)-1)=20.5

[4] (0.2251739)*Sentence Length +(-0.0114461)*Sentence Length^2 becomes larger towards Sentence Length=10, and smaller after Sentence Length=11.

[5] (0. 0706634)*Sentence Length +(-0.0023444)*Sentence Length^2 becomes larger towards Sentence Length=15, and smaller after Sentence Length=16.

[6] 100*(exp(-0.394)-1)=-32.6, 100*(exp(-0.245)-1)=-21.7, 100*(exp(-0.022)-1)=-2.2

	I	II	III	IV
Sample:	All sentences	Incomplete sentences	Simple sentences	Complex sentences
Independent variables:				
GTM Score	2.208***	1.811***	2.720***	1.863***
(Range: 0 - 1)	(0.08)	(0.12)	(0.15)	(0.17)
Sentence Length	0.070***	0.225***	0.071***	-0.000
(Range: 1 - 41)	(0.01)	(0.02)	(0.02)	(0.02)
Sentence Length^2	-0.002***	-0.011***	-0.002***	0.000
(Range: 1 - 41)	(0.00)	(0.00)	(0.00)	(0.00)
Dependency Error	-0.224***	-0.394***	-0.022	-0.245***
(0: Absent / 1: Present)	(0.04)	(0.07)	(0.05)	(0.05)
Post-Editor A	-0.926***	-0.714***	-0.927***	-1.255***
	(0.04)	(0.06)	(0.05)	(0.05)
Post-Editor B	0.018	0.102	-0.021	-0.035
	(0.04)	(0.06)	(0.05)	(0.05)
Adjusted R-squared	0.54	0.50	0.63	0.69
Number of cases	1,413	600	429	384

Dependent variable: Logarithm of post-editing speed (Range: -.301 to 5.011, SD: 0.799)
Unstandardised coefficients are shown in bold face, with standard errors in the parenthesis underneath.
Asterisks indicate statistical significance. ***: p < 0.01

Table 4. Multiple regression analysis of post-editing speed by sentence structures

5 Conclusion

This paper first examined the correlation between various automatic metric scores and human PE speed. It was found that different metrics had different types of relationship with PE speed, and the Pearson correlation coefficients ranged from 0.50 to 0.56. Among the tested metrics, GTM showed the highest correlation with PE speed, but the level of correlation differed greatly depending on sentence structures; it was stronger for simple sentences, and weaker for incomplete and complex sentences. This may suggest that GTM scores can be a better estimator of the amount of PE effort for simple sentences than for other sentence structures.

Further analysis was conducted to take into account other possible variables to explain the gap between GTM scores and PE speed. It was found that sentence length has an impact on PE speed; very short or very long sentences seem to slow down the PE process. However, the level of impact again differed depending on sentence structures; incomplete sentences were particularly susceptible to sentence length. Dependency errors also had greater impact on incomplete sentences than on complex sentences, and had little impact on simple sentences.

Overall, it has been suggested that PE speed may not always have a linear relationship with textual differences measured by automatic metric scores, but various conditions including source text characteristics and MT errors may affect PE speed. It is hoped that further investigation into such conditions will grant us a better understanding of the human PE process and give us useful information for improving MT workflow process and providing more effective PE guidelines and training.

Statistical analyses have helped build an organised view of the relationship between MT translation and the human PE process, though the models are yet to be completed. This pilot project involved only three post-editors, which is not sufficient to generalise the findings. A plan for a more extensive experiment with a larger number of post-editors is in progress.

Acknowledgments

This study was supported by Enterprise Ireland and Symantec Corporation. The author of this paper would like to acknowledge the help received from Dr. Hollowood and his team during the course of this project.

References

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2008) Further Meta-Evaluation of Machine Translation. In: *Proceedings of The Third Workshop on Statistical Machine Translation*, Columbus, Ohio, USAAssociation for Computational Linguistics, , pp. 70-106.

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*, San Diego, CA, pp. 138-145.

Guerberof, A.A. (2008) Post-editing MT and TM: a Spanish case. *MultiLingual*, September, 45-50.

Krings, H.P. (2001) *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* The Kent State University Press, Kent, Ohio.

Melamed, I.D., Green, R. & Turian, J.P. (2003) Precision and Recall of Machine Translation. In: *Proceedings of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 61-63.

O'Brien, S. (2006a) EYE-TRACKING AND TRANSLATION MEMORY MATCHES. *Perspectives Studies in Translatology*, 14, 3, 185-205.

O'Brien, S. (2006b) *Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*, DCU.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 311-318.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.

TAUS (2009) *TAUS MARKET ANALYSIS: The Innovation and Interoperability Roadmap for the Translation Industry.* Translation Automation User Society.

Turian, J.P., Shen, L. & Melamed, I.D. (2003) Evaluation of Machine Translation and its Evaluation. In: *Proceedings of MT Summit IX*, New Orleans, USA, pp. 386-393.