ISSN 2007-9737

Statistical Error Analysis of Machine Translation: The Case of Arabic

Mohamed El Marouani, Tarik Boudaa, Nourddine Enneya

Ibn-Tofail University, Faculty of Sciences,
Laboratory of Informatics Systems and Optimization,
Morocco

mohamed.elmarouani@gmail.com, tarikboudaa@yahoo.fr,enneya@uit.ac.ma

Abstract. In this paper, we present a study of an automatic error analysis in the context of machine translation into Arabic. We have created a pipeline tool allowing evaluation of machine translation outputs and identification of errors. A statistical analysis based on cumulative link models is performed also in order to have a global overview about errors of statistical machine translation from English to Arabic, and to investigate the relationship between encountered errors and the human perception of machine translation quality. As expected, this analysis demonstrates that the impact of lexical, semantic and reordering errors is more significant than other errors related to the fluency of the machine translation outputs.

Keywords. Machine translation evaluation, error analysis, cumulative link models, Arabic NLP.

1 Introduction

Handling the quality of machine translation (MT) is a challenging task where the purpose is to determine if a machine translation system responds to the user's requirements. Most metrics are holistic and provide a global score which is based on the comparison of the source and translated sentences. This score is not sufficient to fully identify the weaknesses of an MT system. Thus, identification and analysis of translation errors is beneficial (1) to know the particular strengths and weaknesses of an MT system, (2) to identify modification types which can improve the quality and (3) to discover if a worse-ranked system outperform a better-ranked one in such aspect. Error analysis can be considered as a diagnosis of MT issues in the process of development of an MT engine or during the comparison of several systems. MT research community has established some competitions and shared tasks where the mission is to examine the improvements in machine translation in general and especially in machine translation quality. Error analysis of MT output is not treated sufficiently and almost the totality of metrics provide only a holistic score of MT quality assessment. The performance of these metrics is verified roughly by the calculation of their correlation with the human judgment using coefficients such as Pearson or Kendall [16].

Error analysis in the context of MT is performed either manually or automatically. In both cases, a taxonomy of errors must be defined with a high or a low granularity. Manual methods are very expensive and time consuming, and pose a problem of inter-annotator agreement. The automatic approach is more efficient and consistent than the manual one [23].

TER [26] is a standard and widely used metric where some extensions are developed in order to support particular languages or to include additional linguistic features. Extending its usage scope to error analysis can bring more insights about the quality issues. For this study we adopted AL-TERp [11] which is an extension of TER and is built especially to support Arabic. Our method is based on the treatment of metric output in order to extract error counts for each sentence, and construct a dataset used thereafter to produce some statistics, and to look finally how those errors are related to the human assessment provided in the studied corpus.

The relation between error profiles and the human judgment is approached as a regression problem using an R implementation of cumulative link models [1] that is a special case of logistic regression.

The second section of this paper provides a background about this task of machine translation error analysis among the presentation of different approaches of error analysis, some works involving Arabic language and AL-TERp machine translation evaluation metric which is the basis of our work. The third section presents the methodology adopted in this study. The fourth section is dedicated to the performed experiments. Finally, we conclude this paper in the fifth section.

2 Background

Error analysis is an important field in Natural Language Processing (NLP) in general and especially in MT. It serves among others as a driver of systems' improvement. First important works in the context of MT are performed manually or using basic metrics such as WER or PER [24], and we observe that recent works focus especially on the comparison of neural MT and statistical MT by analysing and comparing obtained errors [6] [4]. We present in the following subsections some relevant previous works segmenting them on manual analysis, automatic analysis and works involving Arabic language as a study case. Finally, we describe the metric AL-TERp giving that it will be the basis of this study.

2.1 Manual Analysis

In order to compare MT systems or to investigate error types during system development process, error taxonomies were designed in several works with different granularities and focusing on several linguistic phenomena. These taxonomies were used by annotators to determine errors for each translated sentence. After the apparition of the first MT evaluation metrics like BLEU [21] and NIST [9], research community has been oriented to know more about the MT performance by examining obtained errors.

In this context, [30] established a hierarchical classification of translation errors and carried out an analysis following this classification for three language pairs: Spanish to English, English to Spanish and Chinese to English. [13] reported a linguistic error analysis performed over the n-gram based baseline output. The analysis was performed by a Catalan and Spanish native linguist at the level of syntax, semantics and morphology.

[5] examined two techniques of manual evaluation: blind post-editing used in WMT evaluation campaigns and direct explicit marking of errors. This study compares error types for a set of four MT systems translating from English to Czech. Performed analysis leads to an overall conclusion about statistical systems which are better in lexical choice while the fewest morphological errors can be achieved either by a large language model or a deterministic morphological generator.

Other works have built graphical tools to assist human annotators to perform error analysis such as BLAST [28] which aids the user by highlighting similarities with a reference sentence and can be used with any hierarchical error typology.

As mentioned above, human annotation is used recently on the comparison of neural MT systems and statistical MT. These comparisons highlight the strengths and weaknesses of each paradigm. [17] compared three approaches of machine translation (pure phrase-based, factored phrase-based and neural) by performing a fine-grained manual evaluation via error annotation of the systems' outputs. The studied languages pair is English-to-Croatian. The error types in the used annotation are compliant with the Multidimensional Quality Metrics (MQM) error taxonomy [19]. Results demonstrated that the best performing system (neural MT) reduces the errors produced by the worst system (phrase-based MT) by 54%.

[18] extended their first work [17] by performing additional categorisation and analysis of agreement errors and had provided some examples of sentences from the used dataset in this work and more detailed discussions of the obtained results.

Human error analysis is difficult, time consuming and represents in some cases a low inter-annotator agreement. Hence, researchers have been created automatic tools to perform this task.

2.2 Automatic Analysis

On the other hand, automatic analysis of MT errors is carried out by comparing translated sentences to reference ones. This comparison is usually based on a monolingual alignment performed on the most cases as it is done for automatic MT evaluation metrics.

[22] represents a first used tool in this field. It detects five word-level error classes: morphological errors, reordering errors, missing words, extra words and lexical errors. In order to obtain more details, this tool can integrate other information like POS tags. It implements the method based on the standard Word Error Rate (WER) combined with the precision and recall based error rates. It is shown that the obtained results have high correlation with the results obtained by human evaluators.

Addicter [32] is another tool that identifies and labels automatically translation errors. It provides also training and testing corpus browser and word (or phrase) alignment info summarization. This tool relies on the parallel corpora being word-aligned. A light-weight monolingual aligner is included in the tool but is recommended to use an external word aligner such as GIZA++. The translation error taxonomy is taken from [5], which in turn is based on the taxonomy, proposed by [30].

[15] submitted to WMT'12 a human MT evaluation metric called TerrorCat which produces in plus of the global score an automatic error analysis yielding the frequencies of every error category for each translated sentence. This tool uses Hjerson and Addicter as subcomponents, and the features produced by those two tools are used as inputs for a binary SVM classifier trained on the data of manual ranking evaluations of the previous WMT editions.

[2] proposed a new method to perform error analysis task using L1 regularized discriminative language models, and evaluate its effectiveness. Authors concluded that weights trained by discriminative language models are more effective at identifying errors than n-grams chosen either randomly or by error frequency.

To overcome the exploratory analysis of errors, some other studies attempt to examine the impact relations of these errors on the overall quality of machine translation systems. [14] used linear mixed-effects models [3] to build a statistical framework aiming quantification of impact of different error types on MT output quality at the level of human perception and as measured by automatic metrics. This work concerned three translation directions having Chinese, Arabic and Russian as target, and conducted to some important findings such as: the absence of correlation between frequency of errors of a given type and human judgments, and errors having the highest impact can be precisely isolated.

In the same field of MT errors identification, some researchers have worked on the issue of automatic correction of produced errors by applying post-editing treatments. Depfix [25] for instance is one of these tools. It is designed to perform a rule-based correction of errors in English-to-Czech statistical MT outputs.

[29] carried out a multilingual and multifaceted evaluation of NMT and PBMT systems for 9 language directions. This evaluation involved an error analysis via Hjerson with a three word-level error classes (inflexion, reordering and lexical). This study concluded for instance that NMT performs better in terms of inflection and reordering consistently across all language directions.

2.3 Error Analysis for Arabic

In addition to the previously cited study performed by [14] that tackled Arabic among other languages, we find some other works related to error analysis in the context of MT into Arabic. AMEANA [10] is a tool designed to identify morphological errors in the output of a given MT system against a gold reference. AMEANA produces detailed statistics on morphological errors in the output. It also generates an oracularly modified version of the output that can be used to measure the effect of these errors using any evaluation metric. Since AMEANA is language independent, [10] have used it in the study of morphological errors of MT into Arabic.

A project called QALB [31] has been launched in 2013 aiming to build a large corpus of manually corrected Arabic text for building automatic

correction tools for Arabic text. This corpus contained a subset of machine translation outputs annotated by professional annotators in terms of predefined error classification and established guidelines. The works carried out around this corpus and related shared task of automatic text correction for Arabic [20] brought important insights about error analysis especially for texts generated by machine translation systems.

2.4 AL-TERp

AL-TERp is an MT evaluation metric based on TER-plus [27]. This metric which integrates some linguistic features of Arabic realizes a high correlation with human judgments and attempts to handle Arabic specificities. AL-TERp takes into account some flexible matching operations such as stems, synonyms and paraphrases, and generates a detailed output for each reference-hypothesis pair sentence. This output lists among other things edit operations done in order to transform hypothesis sentence into reference one. The edit operations generated at the end of application of a tree-edit distance algorithm are: insertions, deletions, substitutions, shifts, stem matches, synonym matches and phrase substitutions. Then, the edit operations can be viewed as translation errors produced by the MT system. Statistical analysis carried out in this paper is based on the outputs of this metric and attempts to address the relationship between the human judgment of the quality and the data provided by this automatic tool.

3 Approach

In order to employ existent tools in a deep evaluation of MT quality of Arabic, we have picked a variant of a standard metric TER which is errors-oriented and suitable for Arabic, namely AL-TERp. In addition to the global score provided by the metric, we have taken the detailed output explained in the previous section as the raw data of our error analysis module. In the first stage, we have defined a compliant errors taxonomy with five error classes:

1. Missed words (*mis*), which is defined by adding edit operation.
2. Extra words (*ext*), which is defined by deleting edit operation.
3. Lexical choice (*lex*), which is defined by substitution edit operation.
4. Inflection (*inf*), which is defined by stems matching edit operation.
5. Reordering (*reo*), which is defined by shifting edit operation.

We count also synonyms (*syn*) and paraphrases (*par*) edit operations which indicate us using of another style or domain language.

Figure 1 illustrates all error types among two examples that align reference and hypothesis sentences as given by AL-TERp tool.

[11, 12] tackled issues related on how to build a machine translation evaluation tool suitable for Arabic language regarding human judgments, and studying impact of incorporation of linguistic features in the computation of a metric score that correlates well with human evaluation. In the some context, we study in this paper how generated errors by AL-TERp are related to the human judgment. This relation is investigated using a regression model allowing to take into account an ordinal response which is in our case the rank provided by the annotator to each translated sentence.

The used regression model is the cumulative link models. Cumulative link models are a powerful model class for ordinal data, since observations are treated rightfully as categorical, the ordered nature is exploited and the flexible regression framework allows in-depth analyses. We note that while cumulative link models are not the only type of ordinal regression model, they are by far the most popular class of ordinal regression models. Also it is worth also to mention that a cumulative link model can be motivated by assuming an underlying continuous latent variable which is in our case the MT evaluation score [7].

In our case, the cumulative link models can be written as the following:

$$logit(P(R_i \leq j)) = \theta_j + \sum_{k=1}^{8} \beta_k p_k, \qquad (1)$$

Source	China sent 100 tons of relief supplies to Sri Lanka last Wednesday.														
Reference					لسريلانكا	الاغاثة	مواد	من	مائة طن	يوم الاربعاء الماضي			الصين	وقدمت	
Hypothesis	يوم الاربعاء الماضي	سريلانكا	الي		الاغاثة	مواد	من	100 طن			حوالي	ترسل	الصين		
Error type	*reo*		*mis*	*mis*	*ext*	*exa*	*exa*	*exa*	*par*	*reo*	*mis*	*mis*	*exa*	*ext*	

Source	However, the peace plan, drawn up by the United States, United Nations, the European Union and Russia, has not been launched since it was proposed in June 2003.																						
Reference	2003		يونيو	في	طرحها	منذ	تنطلق	لم	وروسيا	الاوربي	والاتحاد	المتحدة	والامم	المتحدة	الولايات	اعدتها	التي	خطة السلام	ان		غير		
Hypothesis	2003	يونيه	حزيران	في	اقترح	حيث			وروسيا	الاوربي	والاتحاد	المتحدة	والامم	المتحدة	الولايات	اعدتها	التي	بخطة السلام	بدا	قد	لا	ذلك	ومع
Error type	*exa*	*mis*	*syn*	*exa*	*lex*	*lex*	*ext*	*ext*	*exa*	*exa*	*exa*	*exa*	*exa*	*exa*	*exa*	*exa*	*exa*	*par*	*lex*	*mis*	*lex*	*mis*	*mis*

Fig. 1. Error types examples

where the $logit$ function is defined as $logit(x) = log(x/1-x)$, R_i is the rank of the i^{th} sentence, j a rank value spanning between 1 and 5, $P(R_i \leq j)$ is the probability that rank of the i^{th} sentence is less than j, θ_j are the intercepts of our model and the β_k are the regression coefficients of each predictor p_k.

For each pair translation and reference sentences, predictors are the normalized counts regarding the reference sentence length of: exact matching words count (*exa*), error count for each already cited error type (*mis*, *ext*, *lex*, *inf*, *reo*) and the remaining edit operations *syn* and *par*.

This regression model is implemented via the R package *ordinal* [8]. The model is fitted by maximum likelihood estimation method, and the used link function is logit function.

We note that our objective in this study is not the prediction of the ordinal output of unobserved data but only the statistical analysis of the relationship between: in one side the ordinal output which is the human ranking of a set of translations provided by five MT systems for each sentence in our dataset, and in other side the explanatory variables, which are the above cited edit operations, extracted for each system and pair translation-reference by our automatic MT evaluation metric AL-TERp.

By treating the relationship between each variable and the human assessment of quality, this detailed study attempts to quantify the importance of each feature introduced in the process of evaluation of MT quality. It is worth also to mention that in contrary of our case the previous works, particularly [14], concern manual annotated errors. As cited, manual annotation of errors is an expensive task and suffers from a consistency problem.

Further to the interpretation of the data given by the performed ordinal regression, other exploratory statistics were reported in this paper in order to give more insights about the distribution of errors by system and by type.

4 Experiments

4.1 Data

This study is based on AL-TERp metric and exploits the same dataset used in related works [11] and [12]. We have adopted the same subset used in the testing phase of the metric AL-TERp. Picking the same dataset allows us to check regarding another point of view the reliability of the previous correlation's studies.

The used dataset is composed of 383 sentences chosen as a test partition in [11] from a dataset of 1383 sentences selected from two corpora: (i) the standard English-Arabic NIST 2005 corpus, commonly used for MT evaluations and composed of political news stories; and (ii) a small dataset of translated Wikipedia articles.

This raw dataset is composed for each entry from the source sentence, the reference one, five translations given by five MT systems and the ranks given by two annotators. The agreement between the two annotators is evaluated to 49.20 in terms of Kendall's τ [12].

Table 1. Edits counts and AL-TERp scores by MT system

	exa	mis	ext	lex	inf	reo	syn	par	AL-TERp
Bing	4316	805	901	1522	778	168	32	418	51.69
CMU	4481	548	1148	1324	692	145	21	360	52.46
Columbia	4433	610	1090	1352	711	154	21	387	52.94
Google	4692	719	822	1369	717	164	22	380	45.58
QCRI	3871	678	1370	1631	732	185	23	385	62.45

Table 2. Parameters of the first cumulative link model

Feature / Error type	Model parameter	Std Error	Pr($>\|z\|$)
exa	-1.796	1.392	0.1972
mis	0.061	0.511	0.9055
ext	2.922	1.419	0.0395*
lex	0.144	1.412	0.9189
reo	3.000	1.288	0.0199*
inf	-1.307	1.442	0.3649
syn	-4.382	3.649	0.2297
par	-1.249	3.045	0.6818

Statistical significance:*** p$<$0.001; ** p$<$0.01; * p$<$0.05

Log Likelihood:-2,992.567

4.2 Experiments and Discussion

Firstly, we evaluate the translations of our dataset produced by the five MT systems using AL-TERp. Those MT systems are three research-oriented phrase-based systems with various morphological and syntactic features (*Columbia*, *QCRI*, *CMU*) and two commercial, off-the-shelf systems (*Bing*, *Google*). According to the experiments of a previous work [11], this evaluation correlates with the human judgments with a Kendall coefficient of 0.3242. In order to introduce our analysis, we expose below some exploratory statistics about the human judgments and the produced errors.

Figure 2 shows rates of ranks assigned for each MT system. Visually, we can observe that *Bing* is the best MT system and *QCRI* is the worst one: for *Bing*, 28% of translated sentences are ranked in the first rank and 26% in the second one. On the other hand, for *QCRI* we have only 6% of sentences in the first rank and 9% in the second one.

As previously mentioned, this metric provides a detailed output as option. We have built a script treating this output in order to extract a set of features for each entry: reference sentence length, count of exact matching words, the counts of each error type and the counts of synonyms and paraphrases matching. Those counts normalized by the reference sentence length, and the rank provided by the human annotator were considered in the next stage as the input of the R package ordinal used to study the relationship between the elements adopted in the calculation of the automatic metric principally the so-called errors and the human perception of quality materialized by the rank assigned by the human annotator.

Table 1 shows the sum of errors by type for each system, and the averaged metric AL-TERp. We recall that in the calculation of this metric, the score is not equally weighted for each edit operation and the weighs were determined by a heuristic optimization process regarding correlation with human annotations [11]. The global correlation coefficients are always under expectation. Then, it is mandatory to examine what types of errors have the highest impact on human perception of quality in order to guide the development of more reliable metrics and to get more convergence.

Table 3. Parameters of selected model

Feature / Error type	Model parameter	Std Error	Pr($>\|z\|$)
exa	-0.912	0.335	0.00645**
ext	3.667	0.503	3.04e-13**
lex	0.983	0.460	0.03265*
reo	3.072	1.288	0.01708*

Statistical significance:*** p<0.001; ** p<0.01; * p<0.05

Log Likelihood:-2,994.076

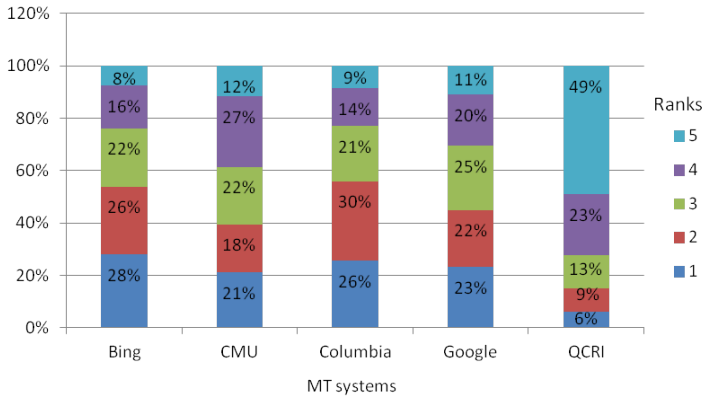

Fig. 2. Systems ranks distributions

Our principal work in this experiment is the study of the relationship between the data summarized in Figure 2 and those of Table 1. We have aimed in the construction of our model to introduce all features given the assumption that they have a significant impact on the model response. First of all, we observe that the introduction of all features in our model don't produce any significant contribution to the response output since the p-value of all features is $>$ to 0.01 (Table 2).

As expected, we observe in Table 2 that features *exa*, *inf*, *syn* and *par* have a negative slopes regarding human judgment ranks (*logit* is a monotonous increasing function). In contrary *mis*, *ext*, *lex* and *reo* have positive slopes. In terms of inference relation, we can conclude that is not judicious to use and combine several features in a metric that will approach MT quality as estimated by humans.

After several experiments, we have selected the subset of coefficients that yields to significant contributions to the model response: *exa*, *ext*, *lex*, *reo*. The results of our experiments that represent the model are reported in Table 3. Features *ext* and *reo* which represent extra words and reordering errors have high impact of respectively 3.667 and 3.072. Substitution edit operation noticed as *lex* error type has an impact of 0.983. Similarly the exact matching feature brings a negative impact of -0.912. Then, we can deduce from these results that the impact of lexical, semantic and reordering errors is more significant than other errors related to the fluency of the machine translation outputs.

5 Conclusions

Learning from errors is an original approach in the building and improvement of systems. Then, identification and analysis of MT errors is really a vector of machine translation enhancement. Studying of MT into Arabic as a challenging direction brought us more informative insights

about issues for this rich morphological language, but focusing on the global correlation between MT evaluation metrics and human quality judgments don't allow us to know how extent correction of a particular error will improve the quality of an MT system.

To the best of our knowledge, the inference method based on cumulative link models is used for the first time for MT error analysis. It is used in studying the relationship between automatically identified errors and human perception of quality. This study evidenced limitations of approaches based on correlation and guided our future works in building more reliable MT evaluation metrics.

In guise of a conclusion, contributions of this work can be summarized in:

— Studying the relationship between human judgments of MT quality and automatic provided features using a new approach and models.

— Investigating the utility of introduction of some features such as those of flexible matching between hypotheses and references.

— Giving more insights about the eventual orientations of research effort in terms of conception and development of MT quality metrics.

References

1. **Agresti, A. (2002).** *Categorical data analysis*. Wiley series in probability and statistics. Wiley-Interscience, New York, 2nd ed edition.
2. **Akabe, K., Neubig, G., Sakti, S., Toda, T., & Nakamura, S. (2014).** Discriminative language models as a tool for machine translation error analysis. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1124–1132.
3. **Baayen, R., Davidson, D., & Bates, D. (2008).** Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, Vol. 59, No. 4, pp. 390–412.
4. **Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2018).** Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. *Computer Speech & Language*, Vol. 49, pp. 52–70.
5. **Bojar, O. (2011).** Analyzing Error Types in English-Czech Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, Vol. 95, No. 1.
6. **Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., & Williams, P. (2017).** A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, Vol. 108, No. 1.
7. **Christensen, R. (2015).** *Cumulative Link Models for Ordinal Regression with the R Package ordinal*.
8. **Christensen, R. (2018).** *ordinal: Regression models for ordinal data. Tech. rep., R package version 2018.8-25*.
9. **Doddington, G. (2002).** Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the second international conference on Human Language Technology Research* -, Association for Computational Linguistics, San Diego, California, pp. 138.
10. **El Kholy, A. & Habash, N. (2011).** Automatic error analysis for morphologically rich languages. *Proc. of the MT Summit XIII, Xiamen, China*, pp. 225–232.
11. **El Marouani, M., Boudaa, T., & Enneya, N. (2017).** AL-TERp: Extended Metric for Machine Translation Evaluation of Arabic. **Frasincar, F., Ittoo, A., Nguyen, L. M., & Métais, E.**, editors, *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, Springer International Publishing, pp. 156–161.
12. **El Marouani, M., Boudaa, T., & Enneya, N. (2018).** Incorporation of Linguistic Features in Machine Translation Evaluation of Arabic. In **Tabii, Y., Lazaar, M., Al Achhab, M., & Enneya, N.**, editors, *Big Data, Cloud and Applications*, volume 872. Springer International Publishing, Cham, pp. 500–511.
13. **Farrús, M., Costa-jussà, M. R., Mariño, J. B., Poch, M., Hernández, A., Henríquez, C., & Fonollosa, J. A. R. (2011).** Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan–Spanish language pair. *Language Resources and Evaluation*, Vol. 45, No. 2, pp. 181–208.
14. **Federico, M., Negri, M., Bentivogli, L., Turchi, M., & Kessler, F.-F. B. (2014).** Assessing the

Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

15. **Fishel, M., Sennrich, R., Popović, M., & Bojar, O. (2012).** Terrorcat: a translation error categorization-based mt quality metric. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Association for Computational Linguistics, pp. 64–70.
16. **Han, L. (2016).** Machine Translation Evaluation Resources and Methods: A Survey. *arXiv:1605.04515v8 [cs]*. ArXiv: 1605.04515v8.
17. **Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2017).** Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, Vol. 108, No. 1.
18. **Klubička, F., Toral, A., & Sánchez-Cartagena, V. M. (2018).** Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian. *Machine Translation*.
19. **Lommel, A., Uszkoreit, H., & Burchardt, A. (2014).** Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica: tecnologies de la traducció*, , No. 12, pp. 455.
20. **Mohit, B., Rozovskaya, A., Habash, N., Zaghouani, W., & Obeid, O. (2014).** The First QALB Shared Task on Automatic Text Correction for Arabic. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 39–47.
21. **Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002).** Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318.
22. **Popović, M. (2011).** Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, Vol. 96, No. -1.
23. **Popović, M. (2018).** Error Classification and Analysis for Machine Translation Quality Assessment. In **Moorkens, J., Castilho, S., Gaspari, F., & Doherty, S.**, editors, *Translation Quality Assessment*, volume 1. Springer International Publishing, Cham, pp. 129–158.
24. **Popović, M. & Ney, H. (2007).** Word error rates: decomposition over Pos classes and applications for error analysis. Association for Computational Linguistics, pp. 48–55.
25. **Rosa, R., Mareček, D., & Dušek, O. (2012).** DEPFIX: A system for automatic correction of Czech MT outputs. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Association for Computational Linguistics, pp. 362–368.
26. **Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006).** A study of translation edit rate with targeted human annotation. *In Proceedings of Association for Machine Translation in the Americas*, pp. 223–231.
27. **Snover, M., Madnani, N., Dorr, B. J., & Schwartz, R. (2009).** Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. *Proceedings of the Fourth Workshop on Statistical Machine Translation - StatMT '09*, Association for Computational Linguistics, Athens, Greece, pp. 259.
28. **Stymne, S. (2011).** Blast: A tool for error analysis of machine translation output. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, Association for Computational Linguistics, pp. 56–61.
29. **Toral, A. & Sánchez-Cartagena, V. M. (2017).** A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. Association for Computational Linguistics, pp. 1063–1073.
30. **Vilar, D., Xu, J., d'Haro, L. F., & Ney, H. (2006).** Error analysis of statistical machine translation output. *Proceedings of LREC*, pp. 697–702.
31. **Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., Farra, N., Alkuhlani, S., & Oflazer, K. (2014).** Large Scale Arabic Error Annotation: Guidelines and Framework. *LREC*, pp. 2362–2369.
32. **Zeman, D., Fishel, M., Berka, J., & Bojar, O. (2011).** Addicter: What Is Wrong with My Translations? *The Prague Bulletin of Mathematical Linguistics*, Vol. 96, No. -1.

Article received on 20/10/2019; accepted on 06/02/2020.
Corresponding author is Mohamed El Marouani.