A Practical Study of Translation Equivalence:

Translation Assessment in a Selected Corpus

Dr. Khaled Bahnasy

Vice Dean of the Faculty of Computer & Information Sciences

Dean of Al Obour High Institute of Management & Informatics

E-mail: khaled.bahnasy@oi.edu.eg

Dr. Tamer Hamed Mohamed

English Linguistics & Translation Lecturer

Department of Humanities, Al Obour High Institutes

E-mail: tamerh@oi.edu.eg

Abstract

The fact that completely programmed translation in numerous sets is away from creating an equivalent or prevalent yield to human translation has led to a serious enthusiasm for translation assessment in the MT people group. Nonetheless, investigation in this field, at this point, has not just overlooked the enormous measures of applicable learning accessible in a firmly related discipline to a great extent, specifically translation ponders; however it also neglected to give a more profound comprehension of the idea of "translation blunders" and "translation quality".

The current paper represents a clarification of the quality notion in translation, by contrasting automatic and human assessments of students' translations in the KOPTE selected corpus. It also demonstrates that studies of translation give modern detailed ideas for translation quality estimation and mistake comment. Moreover, by applying entrenched MT assessment scores, BLUE and Meteor, across KOPTE excerpts which were reviewed by a human professional, the recent research is keen to reveal insight into qualities (possible deficiencies) of such grades.

1- Assessment of quality in translation process

As of late, scientists in MT field assessment have suggested a substantial assortment of strategies in surveying the quality of mechanically delivered translations. Methodologies proceed from completely programmed quality scoring to endeavors aiming the advancement of "human" assessment grades that attempt to misuse the (frequently implicit) linguistic standards of human assessors. The principles as per which quality is evaluated frequently incorporate competence, the level of preservation in meaning, and eloquence, the correctness of target language (Callison-Burch et al., 2007). The objectives of human and the quality of completely programmed assessment are complex and enfold framework enhancement besides investigation and benchmarking.

In translation researches, the scientific (and prescientific) dissertation about the most proficient method to evaluate the quality of human translations continued for a long time. Currently, the improvement of fitting notions and devices has turned out to be significantly more crucial to the field because of the extreme needs of the language activities. Thus, in a completely different conviction, equal to MT, the "quality" of translation could be assessed based on linguistic standard separately, the idea of "translation quality", in specific dissertations, has expected a different structure, separated from a straightforward progress toward equivalence and accepting ideas, for example, functional, pragmatic and stylistic relevance along with text coherence. In this approach, the research gives a review of different methods to deal with translation quality evaluation created in MT and translation researches to indicate how "quality" is defined in both fields and which techniques and properties are utilized. Because of the quantity of writings, the current review is fundamentally inadequate, yet at the same time rational regarding alterations and agreements between MT and human translation assessment.

1.1 Automatic MT quality scores

MT outcomes are generally assessed via independent standards of programmed language which may be utilized to different languages delivered through a MT system. Programmed measurements utilization for MT assessment is appropriate, since MT systems handle a huge amount of data, through which manual assessment would be timewasting and costly.

Automatic Programmed measurements normally process the closeness (adequacy) of a "word" to a "reference" translation and vary from one another by how this closeness in meaning is estimated. The most well-known MT evaluation measurements are IBM BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) that are utilized for modifying MT schemes, as well as assessment measurements for common errands, for example, the Workshop on Statistical Machine Translation (Bojar et al., 2013).

IBM BLEU utilizes n-gram accuracy technique through coordinating programmed translation product against several translations' references. This represents fluency and competence by computing words' exactness, correspondingly the n-gram exactness. Hence, in a way to manage the reliability of normal words, exactness scores are pared, indicating that a word reference is depleted after it is coordinated. Consequently, it is the modified n-gram exactness. For N=4 the n-gram modified exactness is determined and the outcomes are coordinated by utilizing the geometric mean. Rather than reviewing, a briefness penalty is utilized, which reproves that translator products are precise and focused than the reference translations. The NIST metric is a derivative of IBM BLEU. The NIST score is the arithmetic mean of modified n-gram accuracy for N=5 scaled by briefness penalty.

Moreover, NIST considers the amount of data gained of every n-gram, giving more weight to more informative (less recurrent) n-grams and less weight to less informative (more recurrent) n-grams. Meteor is considered another frequently utilized machine translation assessment metric (Denkowski and Lavie, 2011). Unlike IBM BLEU and NIST, Meteor assesses any translator's translation by computing exactness and review on the unigram level also, merging them into a parametrized adjusted mean. The outcome of the adjusted mean is scaled by a discontinuity sanction which penalizes contrasts and gaps in word sequence. Contemporary to these assessment measurements, other measurements are utilized occasionally for the assessment of MT product. The WER (word error-rate) metric according to the Levensthein distance (Levenshtein, 1966), the position independent mistake rate metric PER (Tillmann et al., 1997), and the translation edit rate metric TER (Snover et al., 2006) with its up to date version TERp (Snover et al., 2009) are samples of those metrics.

1.2 Human MT quality assessment

Human assessment of MT product is performed in various ways. The most recurrently existing assessment technique is a naive organization of translated sentences through a "rationale number of evaluators" (Farrús et al., 2010). As indicated by Birch et al. (2013), this type of assessment was utilized, among others, amid the last STATMT workshops and therefore can be more popular. APPRAISE (Federmann, 2012) is a device that can be utilized for that kind of tasks, since it tolerates the manual positioning of sentences, quality estimation, error annotation and post-editing. Different types of assessment exist, for example, Birch et al. (2013) propose HMEANT, an assessment score dependent on MEANT (Lo and Wu, 2011), a semi- programmed MT quality score that estimates the level of meaning maintenance by looking at verb structures and semantic functions of hypothetical translations to their equivalent translations in reference translation(s). Unluckily, Birch et al. (2013) report awkwardness in creating coherent text arrangements between proposed text and its translation, an issue that affects the final HMEANT score computing. However, it appears to be barely amazing regarding the difficulty of comments taken (despite authors' descriptions, some common linguistic perceptions can be expected) and the parameters and organizing are intended to be insignificant.

A secondary human assessment strategy for MT that is utilized for inaccuracy analysis is experiencing reading tests (e.g. Maney et al. (2012), Weiss and Ahrenberg (2012)). In addition, HTER (Snover et al., 2006) is considered a TER-based repair oriented metric which utilizes human illustrators (where the main evident qualificational prerequisite is being fluent in the target language) to create "targeted" reference translations by post-altering the MT product or the current reference translations, following the objective to find the most brief way between the hypothesis and a "equivalent" reference. Snover et al. (2006) report a great relationship between assessment with HTER and conventional human sufficiency and fluency choices. Finally, Somers (2011) makes reference to other repair oriented methods, for example, post-altering exertion estimated by the quantity of key-strokes or time consumed on creating an equivalent translation on the base of MT product.

1.3 Translation Quality

Studies of translation "quality" concentrated basically on equivalence which represents adequacy: effective translation is an ideal agreement between meaning maintenance and target language accuracy, which was particularly applicable to the religious texts' translation. Kußmaul (2000) for example, definitely refers to Martin Luther's Bible translation into German as good translation since Luther, as indicated by his own declaration and following his reformative desire, concentrated on delivering fluent, effectively justifiable content rather than mirroring the linguistic structures of the Hebrew, Aramaic and Greek originals (see additionally Windle and Pym (2011) for further information).

Later work in translation researches has discarded the one-dimensional perspectives of the connection between source and target text and assumes that, according to the communicative context for which and within the translation is created, such a connection can extremely vary. That is, the level of linguistic and semantic "fidelity" of effective translation towards the source text relies upon functional measurements. This notion is represented in the standard of "primary versus secondary", "documentary versus instrumental", and "cover versus overt" translation (Hönig, 2003). Consequently, since in numerous circumstances various translation techniques may be suitably accepted, assessment standard turn out to be basically reliant on the function of the translation role either in language or culture. This opinion is mostly unmistakable through the supposed *skopos* theory (cf. Dizdar (2003)).

At this point, translation mistakes are not only simple basic disruption of the target language framework or absolute breakdowns to translate words or texts, but also disruption of the translation process that can show themselves on all dimensions of text creation (Nord, 2003). Here it is essential to indicate that linguistic errors are only one kind of covered errors as one of the most dominant mistakes in MT divisions, besides stylistic, colloquial, syntactic, linguistic, modal, temporal, cohesive, and other types of mistakes. In addition, the specific errors of translation occur when the translation does not attain its function as a result of pragmatic (ex.

specific forms of text-type), cultural (ex. text regulations, proper nouns, or any other norms), or formal (ex. layout) deficiencies (Nord, 2003). Such errors may represent different weight depending on the appropriate translation technique of a defined translation process.

Moreover, an adjustment in the concept of translation equivalence is managed by the communicative and functional view of translation, which is not sufficiently viewed anymore as the perceptions of "meaning preservation" or "fidelity", however becomes reliant on artistic, connotational, textual, situational, communicative, and intellectual values (any further discussions check Horn-Helf (1999)). In MT assessment, actually, many of those standards are not completely investigated. To wrap things up, the industry of translation has established normative patterns and editing systems. For instance, the DIN EN 15038:200608 (Deutsches Institut für Normung, 2006) considers translation errors, quality management, and qualificational prerequisites for translators and editors, while the SAE J2450 standard (Society of Automotive Engineers, 2005) represents a measured "translation quality measurement". Through an application perception, Mertin (2006) illustrates translation quality management methods in a major automotive organization through which, among different things, builds up a measured translation error system for editing.

1.4 Problem Definition

The previous argument states that the divergences between assessment methodologies improved across the two fields are significant, while the object of assessment is equal for both MT and translation researches. Primarily, in translation studies, assessment is considered a specialist commission where fluency in one or other languages is absolutely insufficient, but where expert specific knowledge is needed. Also in translation studies, another essential variance is that assessment is regularly not performed on the complete sentence level, as sentences may normally break up into numerous "units of translation" also, can definitely include more than a single "translation problem".

Accordingly, the dominant MT system of classifying the entire sentences consistent with some programmed scores, regarding translation studies, may not deliver accurate assessments. Reaching this point, it is clear that the MT group attempt for competence or accuracy in meaning does not fit with the concepts of weighting translation mistakes, implementing various translation techniques, and, thus, does not match the convoluted source/target content relations that have been recognized through translation researches. Assessment strategies that depend on basic approaches of linguistic equivalence, as the n-gram overlap (BLEU) or on the other hand, just a little more convoluted, the protection of syntactic edges and semantic tasks (MEANT) fail to offer clear criteria to differentiate between legitimate and ill-legitimate dissimilarity.

Furthermore, semantic and pragmatic standards besides the concept of "reference translation" stay relatively unclear. Alternatively, the MT group has remarked translation assessment as an uncertain research problem. For instance, Birch et al. (2013) express that organizing decisions is difficult to prevail, while Callison-Burch et al. (2007) perform broad connection trials of an entire range of programmed MT assessment measurements in correlation to human decisions, demonstrating that BLEU does not rank most astounding, but rather still stays in the best section. Despite everything it should be demonstrated how MT research can benefit from more refined assessment measures and whether every single item of the parameters that are viewed as important to the assessment of human translations are pertinent for MT use situations, as well. In the rest of this paper, a study on how much and conceivably for which reasons programmed MT assessment scores (BLEU and Meteor) vary from translation expert quality arbitration on excerpts of a French-German translation student corpus is presented.

2 KOPTE Corpus

2.1 Corpus plan

The KOPTE corpus is a numerous translation learner corpus that includes many translations of the same source text performed by trainee translators. The KOPTE corpus (Wurm, 2013) was planned to facilitate translation assessment research in a college instructional program for translators and to clarify students' translation issues and their problem solving systems. To accomplish this objective, a corpus of students' translations was assembled. The corpus comprises of numerous translations of similar source texts delivered by student translators within a classroom setting. In general, it covers 985 translations of 77 source texts adding up to an aggregate of 318,467 proofs. From French 50 papers, the source texts were taken and converted into German over a range of long time where the translation is prepared to be published in German paper. Thus, translation processes contain the utilization of idioms, clarifications of culture-specific elements, changes in the cohesive components across the whole textual.

2.2 Supervising translation characteristics and its assessment in KOPTE

Students' translations were reviewed by a highly skilled instructor of translation, with the target of offering a practical criticism to students. All translations were reviewed and both mistakes and gains were prominent through the text as per a fine-grained assessment system. Across this system, the influence of any assessed element is shown among numbers varying from plus/minus 1 (minimum) to plus/minus 8 (maximum). Regarding these assessments, each and every translation was given a final review consistent with the German reviewing framework on a scale from 1 ("great") to 6 ("extremely incorrect") with middle dimensions from .0, .3 and .7. To figure out this result, positive and negative assessments were calculated independently, before the negative score was subtracted from the positive one. The score of (around) zero equals to the "great" (=2), and to get "Perfect" (=1) any student requires sufficient positive assessments.

The assessment proposal depends on a strategy that any student translations are evaluated according to extrinsic and intrinsic variables. The extrinsic variables represent the communication where situation is illustrated through the source text and the brief translation (producer, receiver, medium, time, and location). Where intrinsic variables contain (8) classifications: structure, form, stylistics/register, cohesion, lexis/semantics, grammar, specific translation issues, and function. Such classes are holders of more optimized criteria which is applicable to different sections of the (source/target text) or in a way to the entire text, depending on the measurement norms. Some of the internal sub measurements are presented shortly in Table (1). Regarding the KOPTE corpus, the quantitative investigation of mistake types reveals the fact that semantic/lexical mistakes are the most widely recognized mistakes in students' translation (Wurm, 2013). Moreover, in classroom settings, as various assessments are not practical, the assessments in KOPTE are done by only one assessment factor. Despite the fact that various assessments would have been considered very profitable, the information accessible from KOPTE is assessed by an expert academic translator with conventional background in teaching translation.

Besides, the assessment system is substantially more elaborated than the error explanation systems that are typically represented across texts, and it is academically provoked. An examination of the average assessments in the data sample (shown in table (2) and (4)) demonstrates that scoring changes marginally only between various texts, regarding the extreme possible alteration from 1 to 6, and in a way can be represented as being coherent.

Criteria	Examples of Subcriteria
author, recipients, medium, topic, location, time	———
Form Structure Cohesion Stylistic Grammar Semantics Translation problems	Paragraphs, formatting thematic, progression, macrostructure, illustrations, reference, connections, style, genre determiners, modality, syntax, textual semantics, idioms, numbers, terminology, erroneous source text, proper names, culture-specific items, ideology, math units, pragmatics, allusions, goal dependence

function	

Table (1): The Criteria of internal evaluation according to the KOPTE proposal

3 Experiments

In this research, the objective is to consider (in KOPTE) if the human translation judgments can be carried out through the quality measurements of a naive programmed as performed in MT, to be specific BLEU and Meteor. Furthermore, the research aims to:

- investigating how programmed assessment scores affect accurate expert human assessments.
- examining if references higher number enhances the programmed scores or not and why.
- studying if references higher number offers more consistent assessment scores as estimated by an enhanced connection with the human judgments.

Three experiments are conducted here to study the performance of programmed MT assessment scores through operating IBM BLEU (Papineni et al., 2002) in accordance with Meteor (Denkowski and Lavie, 2011) to an excerpt of KOPTE translations that were delivered by translation students planning for their master's final exams. Scores are carried out over the whole texts. Kendall's grade connection coefficient for every text after the system presented in Sachs and Hedderich (2009) is calculated to assess the general performance of the programmed assessment scores on these texts. Connections are considered for:

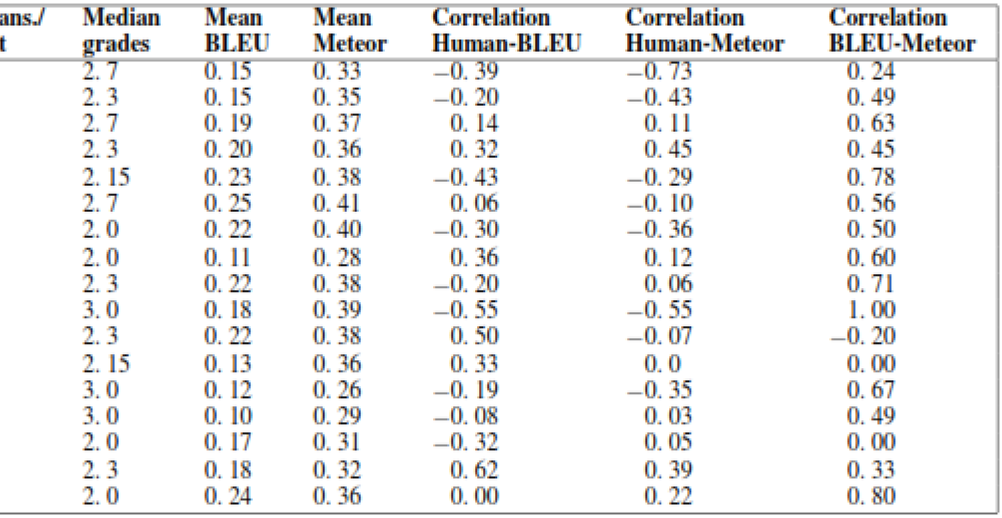
- The human master evaluations and BLEU assessments for every translation.
- The human master evaluations and Meteor assessments for every translation.
- BLEU and Meteor assessments for every translation.

3.1 Experimental framework and results

Within the first experiment, the programmed assessment scores are applied to the source texts in table (2), selecting the student translation, for every text, that gets the best human review as reference translation. Eliminating reference translation, the average human assessments with the mean BLEU and Meteor and associated scores attained for every text are integrated in table (2). In the second experiment, such a strategy is repeated though a set of three reference translations is utilized. Table (3) represents the final results. Eventually, in the last experiment, five reference translations chosen with reference to their human assessment are utilized (see table (4)). In the two stages, source texts with less than four accessible hypotheses are rejected from the data set.

3.2 Discussion

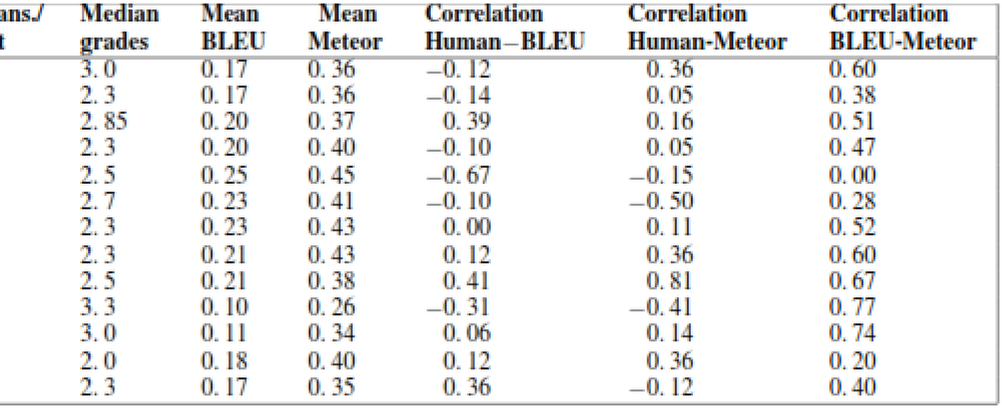
As illustrated in the tables, a set of 152 translations is examined in the first experiment, though in the second and the third experiments such numbers are decreased to 108 and 68 separately as a result of more references selection. The human master assessments appraised the majority of these translations as adequate, as can be shown from the average assessment of each experiment as 2.3 regarding the first one and repeatedly reduced to 3.0 as per the third experiment as a result of the choosing more accurate referential translations.



ans./ t	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
	2. 7	0. 15	0. 33	−0. 39	−0. 73	0. 24
	2. 3	0. 15	0. 35	−0. 20	−0. 43	0. 49
	2. 7	0. 19	0. 37	0. 14	0. 11	0. 63
	2. 3	0. 20	0. 36	0. 32	0. 45	0. 45
	2. 15	0. 23	0. 38	−0. 43	−0. 29	0. 78
	2. 7	0. 25	0. 41	0. 06	−0. 10	0. 56
	2. 0	0. 22	0. 40	−0. 30	−0. 36	0. 50
	2. 0	0. 11	0. 28	0. 36	0. 12	0. 60
	2. 3	0. 22	0. 38	−0. 20	0. 06	0. 71
	3. 0	0. 18	0. 39	−0. 55	−0. 55	1. 00
	2. 3	0. 22	0. 38	0. 50	−0. 07	−0. 20
	2. 15	0. 13	0. 36	0. 33	0. 0	0. 00
	3. 0	0. 12	0. 26	−0. 19	−0. 35	0. 67
	3. 0	0. 10	0. 29	−0. 08	0. 03	0. 49
	2. 0	0. 17	0. 31	−0. 32	0. 05	0. 00
	2. 3	0. 18	0. 32	0. 62	0. 39	0. 33
	2. 0	0. 24	0. 36	0. 00	0. 22	0. 80

Table (2): The source texts, the number of human translations of each source text, average of the acquired assessment of each source text, average of the BLEU and Meteor grades of each source text, and Kendall’s grade connection coefficients of the first experiment

.



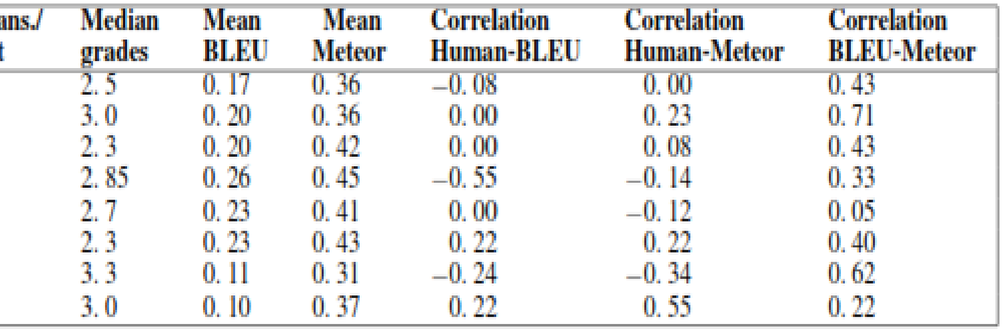
ans./ t	Median grades	Mean BLEU	Mean Meteor	Correlation Human – BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
	3. 0	0. 17	0. 36	–0. 12	0. 36	0. 60
	2. 3	0. 17	0. 36	–0. 14	0. 05	0. 38
	2. 85	0. 20	0. 37	0. 39	0. 16	0. 51
	2. 3	0. 20	0. 40	–0. 10	0. 05	0. 47
	2. 5	0. 25	0. 45	–0. 67	–0. 15	0. 00
	2. 7	0. 23	0. 41	–0. 10	–0. 50	0. 28
	2. 3	0. 23	0. 43	0. 00	0. 11	0. 52
	2. 3	0. 21	0. 43	0. 12	0. 36	0. 60
	2. 5	0. 21	0. 38	0. 41	0. 81	0. 67
	3. 3	0. 10	0. 26	–0. 31	–0. 41	0. 77
	3. 0	0. 11	0. 34	0. 06	0. 14	0. 74
	2. 0	0. 18	0. 40	0. 12	0. 36	0. 20
	2. 3	0. 17	0. 35	0. 36	–0. 12	0. 40

Table (3): The source texts, number of human translations of each source text, average of the acquired assessment of each source text, average of the BLEU and Meteor grades of each source text, and Kendall’s grade connection coefficients of the second experiment.

The best translations’ grades indicated as references across the first and second investigation are in the range of 1.0 and 2.3, though for the third investigation the chosen references are assessed with grades in the range of 1.0 and 2.7. In any case, the average grade for the whole references in the three investigations is constantly 1.7. It is clear that from the general average grade and the indicated translations as references average grade the translations chosen as references are absolutely better than the other ones. For each source text, the BLEU and Meteor scores shown in the previous tables are of poor qualities across the grades of the individual translations. Such grades are very depressing, with a greatest grade of 0.25 across the three investigations for BLEU and 0.45 for Meteor. Though, the translations can't be viewed as meaningless provided the human assessments. Actually, the connection coefficients reveal that both BLEU and Meteor (unless some remarkable instances) do not correspond with the quality of the human assessments, yet, they represent a weak leaning to correspond with one another. In addition, the data demonstrates that adding reference translations did not improve in both considerably developed BLEU or Meteor grades and enhanced connections.

3.3 Qualitative examination

Regarding the fact that human quality decisions are not connected with programmed grades if the purpose of assessment is a human translation (rather than a machine) coordinates prior outcomes displayed by Doddington (2002) across the framework of assessing NIST. Doddington (2002) denotes that "differences between professional translators are far more subtle [than differences between machine-produced translations, the authors] and thus less well characterized by N-gram measurements."



ans./ t	Median grades	Mean BLEU	Mean Meteor	Correlation Human-BLEU	Correlation Human-Meteor	Correlation BLEU-Meteor
	2.5	0.17	0.36	–0.08	0.00	0.43
	3.0	0.20	0.36	0.00	0.23	0.71
	2.3	0.20	0.42	0.00	0.08	0.43
	2.85	0.26	0.45	–0.55	–0.14	0.33
	2.7	0.23	0.41	0.00	–0.12	0.05
	2.3	0.23	0.43	0.22	0.22	0.40
	3.3	0.11	0.31	–0.24	–0.34	0.62
	3.0	0.10	0.37	0.22	0.55	0.22

Table (4): The source texts, the number of human translations of each source text, average of the acquired assessment of each source text, average of the BLEU and Meteor grades of each source text, and Kendall's grade connection coefficients of the third experiment

The research paper presents a qualitative examination of selected translations of KOPTE corpus to investigate if the distinctions between human translations are absolutely as accurate as proposed by Doddington and at least to realize theories that could clarify the poor results of the programmed grades. Three source texts, utilized in the second investigation, are selected here in particular AT008, AT023 and AT053 and check their own reference translations to the chosen translations. This investigation was operated on the lexical level only, which means, a large amount of the KOPTE's features expanded assessment system are definitely not evaluated. However, the investigation indicates that the range of variety that exists only on the lexical dimension is relatively great.

One of the most familiar traits is the occurrence of some naive difference as a result of using equivalent words or employing phrasal variations or rephrasing sentences. Furthermore, the recorded samples reveal that lexical variety can be activated by various source text elements. The observations appeared in the tables are notable translation issues, e.g. proper nouns, conversational or figurative discourse or numbers. Alternate classes in the table are less obvious, that is, they can intersect. In this investigation, components of the source content that can't be translated accurately rather require an innovative arrangement are represented as translation issues.

Distinctive translation techniques can be employed to various types of problems, to the culture-specific elements translation, proper nouns, other source text components, or culture-specific contentions. The specific table in addition to different examples examined here demonstrate that for this classification a few translators included extra data, to adjust the assessment to the German target culture (adjusting pronouns or deictic items) or to adjust the organizing decisions to the variation desired by the target culture (e.g. commas rather than fullstops, and using quotes differently), while other translators select the literally translation. The two techniques are real in specific situations, in any case, it tends to be accepted that adjustments require more prominent intellectual efforts. The ambiguities of the source, regarding the fundamental classification, are features of the source text that can be translated in various styles - nonetheless for a translator processing texts from a different language. Clearly, the link between this type and translation errors isn't sufficiently clear.

Though, it should be noted that for different categories all translations are not similarly impressive - while numerous variations are acceptable and genuine. The best solutions for given errors are scattered unequally through the translations examined. Away from the absolutely lexical features, broad variety can be seen on the syntactic ones, besides the grammatical features. For instance, translators in a way decide to breakdown the complex structure of the French text into less difficult and meaningful sentences, making significant shifts in the text data structure.

Regarding the performance of the programmed scores, the fundamental investigation - that is insufficiently in scale and verification - recommends that neither BLEU nor Meteor can adapt with the amount of varieties found in data. Explicitly, they can't recognize genuine and ill-conceived varieties or specifically maintain trivial issues, however appear to be incapable to coordinate satisfactory variations as a result of lexical and phrasal variety or dissimilar grammatical structures in various verb structures, word arrangements and text segments, far from any acceptable varieties on advanced linguistic categories. This way, programmed scores appear to misrepresent surface dissimilarities and therefore allocate depleted scores to a number of translations that were said to be adequate by a human expert translator.

Regarding the effect of those findings for MT assessment purposes, it isn't clear to expect that the distinctions that exist between the human translations are further subtle (as being irrelevant) than the ones created by programmed translation techniques. On the other hand, the examination proposes that effective translations are featured with imaginative arrangements that are not fairly reproducible but in a way service to accomplish target language clarity and credibility. This is a crucial side of translation quality away from its creation style.

When the context transfers from human translation to programmed translation, it is difficult to perceive any reason why some of the variations that are observed in human translations chosen from KOPTE should sustain acceptable in one way and inacceptable in another, considering the test data and training utilized for improving the MT context: as a result of the appropriate utilization of the creative and productive powers of natural language which should be imitated by MT outcomes, a wide range of varieties exist in human translations.

4- Conclusion

In the current research, the execution of two completely programmed MT assessment measurements have been investigated, to be specific BLEU and Meteor, in contrast with human translation specialized assessments on an example of a translation student across the KOPTE corpus. The programmed scores are investigated in three experiments with a wide number of references and their execution was contrasted with the human assessments by methods of Kendall's rank connection coefficient. It is suggested, through experiments, that both BLEU and Meteor devaluate the accuracy of the translation examined methodically, which means they give the human specialized translations scores that appear to be very low than. Furthermore, they don't reliably associate with the human specialized assessments. A clarification of this breakdown is not clear enough, yet, the consequences of this qualitative investigation propose that scores of lexical comparability are not capable to adapt sufficiently with both standard lexical varieties (rephrasing…) and variations that can be tackled to the specific purpose of translation.

For Meteor, though such inadequacy may be improved by the arrangement of finer alternative and rephrasing choices, the quantity of clear varieties is enormous. Actually, it appears that many reference translations is required to enfold the entire scope of real variations utilized to translate a specific source text, which appears to be less practical! Thus, in what manner, the scores of BLEU or Meteor can be translated when they exist in MT articles? Regarding the current investigations, it appears to be clear that such scores depend on data-driven concept of translation quality, where the level of consistency of the proposed translation with a reference set is measured, which in a way seems to be difficult because those studies established on sets of various reference can't be associated, both BLEU and Meteor scores cannot be comprehensive in different fields. Moreover, the scores of BLEU or Meteor can't be utilized to assess the concept of data-independent quality or even the easiness of translation for target audience which, as presented, relies on a bunch of elements than the mere surface of lexical similarity. Nonetheless, the recent study prompts some inquiries. One of these inquiries is whether the programmed assessment scores can sustain to be utilized for more untreated differences: to recognize "meaningless" translations from "accurate" ones. The refinements introduced by the KOPTE assessment across acceptable translations hinder any attempt to reply such an inquiry. It is propsed that future work may experience a comparison of the MT systems’ mistakes to human translations in a way to answer the question of how or which translation specific norms can be utilized to the assessment of MT systems.

List of Terms

- KOPTE is a corpus and a research project covering a huge variety of thematic approaches since 2009 to offers several aspects to Translation Studies, esp. to research on translation evaluation and translation competence (acquisition).
- MT stands for machine translation
- Meteor automatic evaluation metric scores machine translation hypotheses by aligning them to one or more reference translations to address some of the deficiencies inherent in the BLEU metric, based on exact, stem, synonym, and paraphrase matches between words and phrases. METEOR also includes some other features not found in other metrics, such as synonymy matching, where instead of matching only on the exact word form; the metric also matches on synonyms. For example, the word "good" in the reference rendering as "well" in the translation counts as a match.
- BLEU, one of the most popular in the field of translation assessment, is one of the first metrics to report high correlation with human judgments of quality. Using an n-gram method, the metric calculates scores for individual segments, generally sentences—then averages these scores over the whole corpus for a final score.
- IBM BLUE is presented by IBM scientists. It is a method for the automatic evaluation of machine translation (MT), the Bilingual Evaluation Understudy; or simply, BLEU.
- N-gram method is a probabilistic language model often used in computational linguistics.
- NIST metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST calculates how informative a particular n-gram is: when a correct n-gram is found, the rarer that n-gram is, the more weight it is giv-

en. NIST also differs from BLEU in its calculation of the brevity penalty.

- Word error rate (WER) is a metric of evaluating machine translation that is based on the Levenshtein distance, and is used for measuring the performance of speech recognition systems. While the Levenshtein distance works at the character level, WER works at the word level. The metric is based on the calculation of the number of words that differ between a piece of machine translated text and a reference translation.
- HMEANT metric is a semi-automatic evaluation metric used for scoring translation quality by matching semantic role fillers on the Polish language.
- STAME 'Statistical Machine Translation'
- HTER "Human-targeted Translation Error Rate" is an edit-distance measure: "the fewest modifications (edits) required to the system output so that it captures the complete meaning of the reference, using relatively fluent English". The formal definition is: HTER = (Substitutions + Insertions + Deletions + Shifts)/Reference Words
- TER-based 'Translation Edit Rate' is an automatic evaluation metric of machine translation (MT). It is an evaluation metric and alignment tool that addresses several of translation weaknesses through the use of paraphrases, stemming, synonyms, as well as edit costs that can be automatically optimized to correlate better with various types of human judgments.
- DIN EN is a certification issued by the German institute for standardization to certify personnel working in Non-destructive testing. This standard evaluates and documents the competence of personnel whose tasks require knowledge of non-destructive tests.
- EN 15038 is a quality standard developed especially for translation services providers. The EN 15038 standard ensures the consistent quality of the service. It requires regular audits by

the certification body, and if any discrepancy is found, the certification shall be revoked.

References

Alexandra Birch, Barry Haddow, Ulrich Germann, Maria Nadejde, Christian Buck, and Philipp Koehn. 2013. The feasibility of HMEANT as a human MT evaluation metric. In Proceedings of the 8th Workshop on SMT, pages 52–61.

Andrea Wurm. 2013. Eigennamen und Realia in einem Korpus studentischer Übersetzungen
(KOPTE). trans-kom, 6(2):381–419.

Brigitte Horn-Helf. 1999. Technisches Übersetzen in Theorie und Praxis. Franke.

Chi-Kiu Lo and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In Proceedings of the 49th Annual Meeting of the ACL, pages 220–229.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In Proceedings of the 2nd Workshop on SMT, pages 136–158.

Christian Federmann. 2012. Appraise: An opensource toolkit for manual evaluation of machine translation output. PBML, 98:25–35, 9.

Christiane Nord. 2003. Transparenz der Korrektur. In Handbuch Translation, pages 384–387. Stauffenburg.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In Proceedings of the EUROSPEECH, pages 2667–2670.

Deutsches Institut für Normung. 2006. DIN EN 15038:2006-08: Übersetzungsdienstleistungen-
Dienstleistungsanforderungen. Beuth.

Dilek Dizdar. 2003. Skopostheorie. In Handbuch Translation, pages 104–107. Stauffenburg.

Elvira Mertin. 2006. Prozessorientiertes Qualitäts management im Dienstleistungsbereich Übersetzen. Peter Lang.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the 2nd International Conference on HLT, pages 138–145.

Hans Hönig. 2003. Humanübersetzung (therapeutisch vs. diagnostisch). In Handbuch Translation, pages 378–381. Stauffenburg.

Harold Somers. 2011. Machine translation: History, development, and limitations. In The Oxford Handbook of Translation Studies, pages 427–440. Oxford university Press.

Kevin Windle and Anthony Pym. 2011. European thinking on secular translation. In The Oxford Handbook of Translation Studies, pages 7–22. Oxford University Press.

Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the ACL, pages 311–318.

Lothar Sachs and Jürgen Hedderich. 2009. Angewandte Statistik. Methodensammlung mit R. Springer.

Maja Popovi´c, Eleftherios Avramidis, Aljoscha Burchardt, Sabine Hunsicker, Sven Schmeier, Cindy Tscherwinka, David Vilar, and Hans Uszkoreit. 2013. Learning from human judgements of machine translation output. In MT Summit, pages 231–238.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the 6th Workshop on SMT, pages 85–91.

Mireia Farrús, Marta R. Costa-Jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguisticbased evaluation criteria to identify statistical machine translation errors. In Proceedings of the 14th Annual Conference of the EAMT, pages 167–173.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of AMTA, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In Proceedings of the 4th Workshop on SMT, pages 259–268.

Ondˇrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. Proceedings of the 8th Workshop on SMT. ACL.

Paul Kußmaul. 2000. Kreatives Übersetzen. Stauffenburg.

Sandra Weiss and Lars Ahrenberg. 2012. Error profiling for evaluation of machine-translated text: a polish-english case study. In Proceedings of the Eighth LREC, pages 1764–1770.

Sara Stymne and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. In Proceedings of the 8th LREC, pages 1785–1790.

Society of Automotive Engineers. 2005. SAE J2450:2005-08: Translation Quality Metric. SAE.

Tucker Maney, Linda Sibert, Dennis Perzanowski, Kalyan Gupta, and Astrid Schmidt-Nielsen. 2012. Toward determining the comprehensibility of machine translations. In Proceedings of the 1st PITR, pages 1–7.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 10(8):707–710.