Out of the *BLEU*: An Error Analysis of Statistical and Neural Machine Translation of WikiHow Articles from English into Arabic

Nessma Diab
Teaching Assistant
Faculty of Al-Alsun (Languages), Ain Shams University
Cairo, Egypt

Abstract

Most studies that compare the quality of Neural Machine Translation (NMT) to that of Statistical Machine Translation (SMT) rely on automatic evaluation methods, mainly the bilingual evaluation understudy (BLEU), without performing any kind of human assessment. While BLEU is a good indicator of the overall performance of MT systems, it does not offer any detailed linguistic insights into the types of errors generated by those MT models. Such insights are crucial for researchers to identify areas for improvement and for language service providers to understand how upgrading to NMT gives them better results. This paper breaks free from BLEU by conducting an error analysis that compares the performance of Google SMT and NMT engines for English-into-Arabic translation. The corpus consists of six WikiHow articles. The analysis is guided by the DQF-MQM Harmonized Error Typology which classifies translation errors into eight major categories, namely, accuracy, fluency, terminology, style, design, locale convention, verity and other (for any other issues). A fine-grained classification of translation errors as such enables the researcher to explore the error types generated by each MT model, the error types eliminated by NMT, and the new error types introduced by NMT. The paper focuses on the English-Arabic language pair because it is one of the least studied pairs in the comparative literature of SMT and NMT. The results show that NMT generates less grammatical errors and mistranslations than SMT. NMT output is more fluent and robust. However, SMT is more consistent with translating proper nouns and out-of-vocabulary words.

Keywords: DQF-MQM harmonized error typology, neural machine translation, statistical machine translation, translation quality assessment

الملخص العربى

تعتمد معظم الدراسات التي تقارن بين جودة الترجمة الآلية العصبية والإحصائية على المقاييس الآلية وبالأخص مقياس "بلو" (BLEU)، وهو خوارزمية تقوم فكرتها على قياس مدى التطابق بين الترجمة الآلية ونموذج لترجمة بشرية، وكلما طابقت الترجمة الآلية تلك البشرية، كان ذلك دليلًا على جودتها. وعلى الرغم أن مقياس "بلو" يُعد مؤشرًا جيدًا على التحسينات التي أُجْرِيَت على نظام الترجمة الآلية بشكلٍ عام، فهو لا يقدم أي معلومات مفيدة عن المشكلات أو الأخطاء التي تولدها تلك النظم؛ لذلك تقوم هذه الدراسة بإجراء تحليل لغوي لأخطاء نظامي الترجمة الآلية العصبية والإحصائية التابعين لشركة جوجل. ويهدف هذا التحليل إلى معرفة الأخطاء التي يولدها كل نظام وتصنيفها، وما هي الأخطاء التي استطاع النظام العصبي تجنبها، وما هي الأخطاء الجديدة التي ارتكبها ولم تكن موجودة في منافسه، بالإضافة إلى الأنماط المسببة لكل خطأ. ويساهم هذا التحليل في تحديد مجالات التطوير ومساعدة مقدمي خدمات الترجمة على اختيار النظام الأفضل لزيادة إنتاجيتهم. وتركز الدراسة على الترجمة من الإنجليزية إلى العربية لأنها من الثنائيات اللغوية الأقل دراسةً في هذا السياق. تتكون عينة البحث من ست مقالات "WikiHow" في مجالات الصحة والأمن السيبراني والعملات الرقمية، ويعتمد التحليل على نموذج "DQF-MQM" لتصنيف الأخطاء والذي يتكون من ثمان فئات رئيسية وهي الدقة، والفصاحة، والمصطلحات، والأسلوب، والتصميم، والاصطلاحات المحلية، والمصداقية، وفئة إضافية لأي أخطاء أخرى. توضح النتائج أن النظام العصبي يولد أخطاءً أقل عددًا وجسامة من النظام الإحصائي بشكلٍ عام، وبالأخص من حيث النحو والتراكيب والمفردات لذا فهو أكثر فصاحة ودقة؛ إلا إنه أقل ثباتًا في الأداء من النظام الإحصائي في ترجمة أسماء الأعلام والكلمات الغريبة عليه.

الكلمات المفتاحية: تصنيف الأخطاء بنموذج DQF-MQM، الترجمة الآلية العصبية، الترجمة الألية الإحصائية، تقييم جودة الترجمة

1. Introduction

Billions of words and media items are posted on the internet every day. Due to the lockdown caused by Covid-19, online content consumption has been doubled and ecommerce industry flourished, with Amazon being the most profit-making company in 2020 with $4 billion ("Prospering in the pandemic", 2020). However, Arabic accounts for only 1.2% of the content languages on the internet. Relying on the human factor alone to translate this gigantic amount of content is neither time nor cost-effective. As a result, many companies resort to using Machine Translation (MT), with different levels of pre- and post-editing, to speed up the translation process and reduce expenses.

Statistical Machine Translation (SMT) was the dominant MT model, until the application of deep learning techniques in MT, in what has become known as Neural Machine Translation (NMT). By the mid-2010s, NMT gained more attention from MT researchers; and by 2016, "the entire research field went neural" (Koehn, 2020, pp. 39-40). NMT was propagated for as the silver bullet that would solve all issues of SMT. Some even claimed that NMT systems achieved human parity (Hassan et al.,2016). However, the human parity hyperbole came under fierce criticism from the research community (Laubli et al., 2018; Toral et al.,2018).

Since 2016, many studies have compared the performance of NMT to that of its statistical predecessor claim that NMT outperforms SMT. To mention a few, Wu et al. (2016) report that deploying the neural model in Google Translate reduces errors by an average of 60% compared to the statistical phrase-based model for English-French, English-Spanish, and English-Chinese language pairs. After comparing the performance of NMT and SMT across the thirty translation directions of the UN Parallel Corpus v1.0, Junczys-Dowmunt et al. (2016) conclude that "for all translation directions NMT is either on par with or surpasses phrase-based SMT" (p. 7). Almahairi et al. (2016), Durrani et al. (2017) and Alrajeh (2018) agree that NMT achieves higher BLEU scores than SMT for English-Arabic translation. Their datasets consist of news articles, TED talks, and the UN corpus.

BLEU is an automatic evaluation metric which produces a numeric value that represents the similarities between the output of the MT engine and a reference translation produced by professional human translators (known as the 'Gold References'). While BLEU is a handy evaluation metric, it fails to provide a detailed performance diagnosis of a given MT model.

Using BLEU scores to compare the performance of NMT and SMT does not answer questions about the type of errors each model generates, the type of errors eliminated in NMT, and the type of new errors introduced by NMT. That is one reason why Kohen (2020, p. 60) describes BLEU scores as "meaningless", i.e., no one knows what the score means. For example, does a score of 0.4 mean that the MT is good enough to translate movie subtitles? Does it mean it is bad for medical texts? Answers to these questions require manual inspection of the output of each MT model, guided by a fine-grained typology of translation errors as the one carried out in this study.

This study compares the two models of Google MT engine: the old statistical model and the new neural one. The comparison relies on English-into-Arabic translation of a corpus of six articles (6,641 words) about cybersecurity, cryptocurrency, and healthcare collected from the WikiHow website. After having the articles translated separately by each MT model, all errors in the output of each model were manually annotated in accordance with the fine-grained DQF-MQM Harmonized Error Typology (Lommel, 2018). It is a shared industry standard that is used to classify and count translation errors sentence-by-sentence according to eight main categories and 33 subcategories and four severity levels.

Annotation results show that NMT outperforms SMT, indeed. NMT generates less grammatical errors and mistranslated words (i.e., the translated content does not accurately reflect the meaning of the source text). NMT output is more fluent, where fluency means a structurally correct text that has no grammatical, spelling or punctuation mistakes. NMT is more robust since the number of untranslated words in SMT output is double the number in NMT output. NMT is more loyal to source texts in that addition errors in the neural output are way less than those in the statistical output, 4 compared to 64 errors, respectively. However, NMT is less consistent than SMT regarding word choice; that is to say, the same word may be translated differently, transliterated, or left untranslated.

2. Research Questions (RQs)

This study attempts to answer the following questions:

1. Which MT model produces fewer errors?
2. Which MT model produces less severe errors?
3. What are the types of errors produced by each system?
4. Which error types are eliminated by NMT?
5. What types of new errors does NMT introduce?

6. What are the patterns that trigger these errors?

3. Significance of the Study

The contribution of this study is threefold. First, it uses an industry-wide standard evaluation model with deep linguistic insights to compare the performance SMT and NMT against the same corpus. Second, it focuses on English-into-Arabic translation of user-generated content, which is an understudied area in the literature comparing NMT to SMT. Finally, the study shows the importance of manual translation assessment for both researchers and language service providers (LSPs). A detailed understanding of MT output tells researchers which areas to improve and helps LSPs know which areas they will improve when they upgrade to NMT.

4.Background

4.1. Automatic Translation Evaluation Metrics

One of the most famous automatic metrics for MT is BLEU (Papineni et al., 2002). It was developed by IBM and stands for bilingual evaluation understudy. BLEU measures how close the output of the MT system is to "the gold standard" (i.e., reference human translations) and produces a score to represent the similarity. The closer the MT output to any reference translation, the higher the score BLEU rewards it. (Kohen, 2020, pp. 53-54).

Because BLEU is fast and inexpensive, it has been used to evaluate several MT models including rule-based models (Simard et al., 2007), statistical models (Dreyer et al., 2007), and neural models (Wu et al., 2016). Moreover, BLEU has been the "de facto" standard for MT evaluation research (Castilho et al., 2018). In Marie et al. (2021), the authors annotated 769 research papers in the field of MT evaluation from 2010 to 2020 based on the evaluation methods used. They found that almost 99% of these papers use BLEU scores to evaluate MT quality, and that 74.3% rely exclusively on BLEU scores without using other automatic metrics and performing statistical testing or referring to human evaluation to ensure that the results are not coincidental.

BLEU also gained popularity because it has been shown to correlate with human judgement. However, there are cases where BLEU fails to correlate with human judgement. For instance, Charniak et al. (2003) used BLEU and human raters to evaluate three MT systems. They reported quality improvements according to human judgements that were given poor BLEU scores. Others similar cases where the BLEU scores did not agree with human evaluations include Callison-Burch et al. (2006), Koehn and Monz (2006), Callison-Burch et al. (2007). This is

because BLEU is biased to reward local matches over the overall accuracy since it is based on exact lexicon matches; therefore, it tends to favor SMT over other MT systems (for instance rule-based MT where other valid variants are used). Therefore, it is inappropriate to use BLEU scores to compare MT systems which use different approaches (Koehn, 2020).

BLEU has also been criticized for being a similarity measure, rather than a quality measure. BLEU's assumption is that if a given translation matches another good translation, then it is equally good by extension. Such similarity is measured based on overlapping n-grams. N-grams are any chunks of texts consisting of two or more consecutive words and which are not linguistically motivated (Sampson, 2003). This means that it does not care about the completeness of meaning or the grammaticality of the sentence and will give a high score to any matching sequence of words, coherent or not.

While a high BLEU score may be indicative of good quality, a low score does not always mean that the translation quality is poor (Culy & Riehemann., 2003). Even a high BLEU score is not always a guarantee of good quality as table 1 (adapted from Linares, 2008, p. 31) demonstrates. In this case, the highest score is given to the most meaningless candidate B only because it has the highest number of 4-gram matches: *right in front of*, *in front of the*, and *front of the lake*—compared to only 1 in candidate A and 0 in candidate C. Consequently, the metric gives segments where no 4-gram matches are found in the reference translation a zero, regrdless of their lower n-gram matches. Therefore, segements that are originally less than 4-grams will always be given a zero (Stroppa et al., 2007). This is because BLEU is designed to assess quality on a system level, not on a segment level. But this brings about another shortcoming: human evaluations of adequacy and fluency or engine rankings are often performed segment by segment; thus, BLEU's correlation with these evaluations will be very low.

Table 1

An Example of a Meaningless Sentence Given the Highest BLEU Score

Source Text	كان البيت الأخضر يطل على البحيرة مباشرة .	
Reference	The green house was righ in front of the lake .	
		BLEU
Output A	The green house was by the lake shore .	0.30
Output B	The green potato right in front of the lake was right .	0.52
Output C	A green house was by the lake shore .	0.00

One of BLEU's major drawbacks is its inability to reward near matches such as synonyms and morphological variations (Koehn, 2020). For example, if the reference translation has the word *beautiful*, but the MT output has the word *pretty*, BLEU will not give any credit to this semantic alternative and will even penalize the MT system for using it. The same applies to morphological differences. According to the metric, there is no similarity whatsoever between *love* and *loves.*

The metric also does not take into account the severity of the errors made by the MT system; it treats all word the same way regardless of their importance to meaning. For instance, using the article "a" instead of "an" before a word that starts with a vowel is not as severe as a missing negation or a content-bearing word; yet such severities are irrelevant to BLEU and do not entail additional penalties (Koehn, 2020).

4.2. Fine-Grained Linguistic Models for Translation Quality Assessment

While manual assessment of MT quality is tedious and time consuming, it provides deep linguistic insights into the type of errors generated by each MT model. One of the most comprehensive error typologies is the DQF-MQM Harmonized Error Typology (Lommel, 2018).

The Multidimensional Quality Metrics (MQM) and the Dynamic Quality Framework (DQF) started as two separate projects. The MQM was developed in 2012 by the German Research Centre for Artificial Intelligence (DFKI). The final version of MQM included a very comprehensive range of 182 error types. However, such granularity never meant that the developers advocated the adoption of all described issues in one evaluation task. Instead, MQM was designed with flexibility in mind. Evaluators can use any subset of errors they deem fit for the purpose and type of the translation under assessment.

In the same year, the Translation Automation User Society (TAUS) developed the DQF, which started out as a mere error typology but was later upgraded to an analytics platform that encompasses several evaluation methods including a content profiler, productivity tests, adequacy and fluency tests, engine rankings and recently a quality dashboard. The original error typology contained six main categories that were based on the issues commonly reported by LSPs. These categories were defined and further divided into specific subgroups. The typology, however, was very similar to some of the issues described in MQM.

For funding-driven reasons and in order to avoid confusion among the users of both typologies, the two joined efforts in 2015 and developed the DQF-MQM harmonized error typology. It is now a widely used industry standard for evaluating both human and machine translation errors. As of

2018, the typology has been used in the standardization process at ASTM International for quality assurance in translation. It is also used by Translators Without Borders (TWB), a non-profit organization, to evaluate the quality of its crowd-sourced volunteer translation in the humanitarian field.

The DQF-MQM harmonized typology presents a formal predefined rubric that makes the evaluation process as consistent and subjective as possible. The typology is integrated into the DQF Quality Dashboard, which has the benefit of tracing the errors back to a specific location in the translated text digitally, instead of just referring to them in an overall feedback on the translation; thus, providing MT developers with error-annotated bilingual corpora. The typology is also compatible with many translation management systems such as Trados Studio, XTM Cloud, and GlobalLink through the DQF plugin, which facilitates the annotation process. It is also available on ACCOLÉ[1] (Esperança-Rodier et al., 2019), a collaborative platform of error annotation for aligned corpora.

The DQF-MQM harmonized typology classifies errors into eight main categories (accuracy, fluency, terminology, style, design, locale convention, verity and other) and 33 subcategories. The most relevant to the present study are defined in table 2 (TAUS, n.d.) [2]. These categories can be customized to fit the purpose of the translation, i.e., some categories can be neglected if they are irrelevant to the text at hand. For example, if the text does not include any formatting, the "design" category can be dropped from the evaluation process.

Table 2

A Subset of DQF-MQM Harmonized Error Typology

Main Category	Subcategory	Definition
Accuracy	Addition	The target includes information not present in the source, for example, adding a date that does not exist in the source text to the translation
	Omission	Content is missing from the translation that is present in the source, for instance, deleting the negation in the translation
	Mistranslation	The target content does not accurately represent the source content, for instance, translating "Apple" brand into تفاحة

[1] http://lig-accole.imag.fr/app.php/login

[2] For a full list of definitions for each error category and subcategory, visit https://www.taus.net/qt21-project#harmonized-error-typology

Main Category	Subcategory	Definition
	Untranslated	Content that should have been translated has been left untranslated, for example, leaving the word "allergy" untranslated in the target
Fluency	Spelling	A word is misspelled, for instance, translating the word "glass" as زخاخ instead of زجاج
	Grammar	Issues related to the grammar or syntax of the text such as function words, word order, agreement, tense, and parts of speech. For instance, translating "the red car" into " الحمراء السيارة" is classified under incorrect word order.
	Inconsistency	Same word in the same context is translated differently, for example, translating the word "vaccine" once as لقاح and again as طعم.
Style	Awkward	A text is written with many embedded clauses and an excessively wordy style, for instance, translating "your" as الخاص بك instead of using the possessive pronoun in Arabic.
	Unidiomatic	The text is grammatical but unnatural
Other		Any other issues

The DQF-MQM error typology also categorizes errors according to their severity into four levels (critical, major, minor, and neutral) and assigns each severity level a penalty score (10, 5, 1, and 0 points respectively). Critical errors are those which render the translation unusable. They also may carry harmful real-life implications. For example, translating an emergency phone number like 911 into the same number for countries outside North America will prevent the caller from getting the urgent help needed; and adding an extra zero to a product's price might make the seller face legal charges. Similarly, major errors hinder the reader from understanding the meaning of the text but do not cause adverse effects. For example, translating the word "table" as جدول instead of طاولة in the Arabic edition of a furniture catalogue is a major error. Minor errors, on the other hand, make the text awkward yet still fulfilling its purpose. They do not affect the overall meaning. A common mistake such as repeating the word كلما when translating comparative correlatives into Arabic will not affect the meaning and will pass unnoticed by most readers. The neutral level is used to report issues that do not count as errors because they do not affect the meaning or the fluency of the text. These include a reviewer's preferred style or

modifications made to the terminology after the translation has been submitted.

4.3. Comparing NMT to SMT

Many studies compare the performance of NMT to that of SMT for a variety of language pairs and text genres. Toral and Sanchez-Cartagena (2017) carried out a multifaceted automatic evaluation to compare NMT and SMT for news translation in nine languages, not including Arabic. They found out that NMT performs better in terms of fluency and morphological inflection. However, NMT is worse than SMT when translating very long sentences (50+ words).

Popović (2017) performed a manual linguistic error analysis of the neural and statistical outputs for English-German news translation. The error taxonomy she followed was not clearly mentioned in the paper; and according to her results, NMT is better at handling verbs, noun collocations, compounding and articles. However, NMT struggles more with prepositions, ambiguous words, and continuous tenses.

Castilho et al. (2017) evaluated the quality of NMT and SMT for e-commerce, patent, and educational texts from English into four target languages (German, Greek, Portuguese and Russian). In addition to automatic tools, they used different human evaluation methods for each domain. For the e-commerce, they conducted Likert-based surveys to evaluate adequacy, and blind engine ranking where participants ranked the MT systems from best to worst. For the patent domain, they used blind ranking and error annotation based on a taxonomy of seven error types "punctuation, part of speech, omission, addition, wrong terminology, literal translation, and word form" (p.114). For the education domain, the human evaluation was based on measuring the post-editing effort, adequacy and fluency ratings, and a simple error classification that focused on "inflectional morphology, word order, omission, addition, and mistranslation" (p. 116). It is concluded that NMT outperforms SMT based on the automatic measure; however, human evaluation presented mixed results. They also reported that using the neural model improved fluency but the results regarding its adequacy and post-editing effort were inconsistent.

Stasimioti & Sosoni (2019) conducted a comparative study of Google's NMT and SMT engines for English-into-Greek translation of two short texts. They performed a four-level analysis. First, they used automatic evaluation tools, namely BLEU and Word Error Rate (WER) to evaluate quality; their results showed slightly improved score for NMT. Second, the asked evaluators to rate the two outputs for adequacy and

fluency; and the neural model was rated higher for fluency and no part of its output was deemed incomprehensible. Third, they measured the post-editing effort using time tracking, keyboard activity logging, and eye-trackers. They reported that post-editing NMT required slightly less time and keystrokes than SMT. They also noted that the eye-fixation duration was marginally longer for SMT which could indicate that post-editing SMT is more cognitively demanding than post-editing NMT. Finally, they hired two professional translators to flag the errors found in both outputs based on the DQF-MQM error typology. The number of errors was a little higher for SMT. As for the error types, they found that SMT produced more mistranslations, word form and agreement errors than NMT. However, the statistical showed better performance in terminology and did not generate any omissions.

4.4. Error Analysis of NMT and SMT for English-into-Arabic Translation

Studies conducting human evaluation to compare NMT to SMT for English-into-Arabic translation are scarce. More attention is given to evaluating NMT only (Abdelaal & Alazzawie, 2020; Hossain et al., 202). Abdelaal and Alazzawie linguistically analyzed the Google NMT output for translating news articles from Arabic into English. Their results show that omissions and mistranslation of homophonic source words are the most common errors. Hossain et al. (2020) examine NMT performance when dealing with negation. Their datasets consist of news translations. They conclude that negation remains problematic for modern NMT systems.

To the best of the researcher's knowledge, no studies have conducted a fine-grained error analysis as the one performed in this study to compare the quality of NMT to that of SMT.

5. Methods

5.1. Corpus

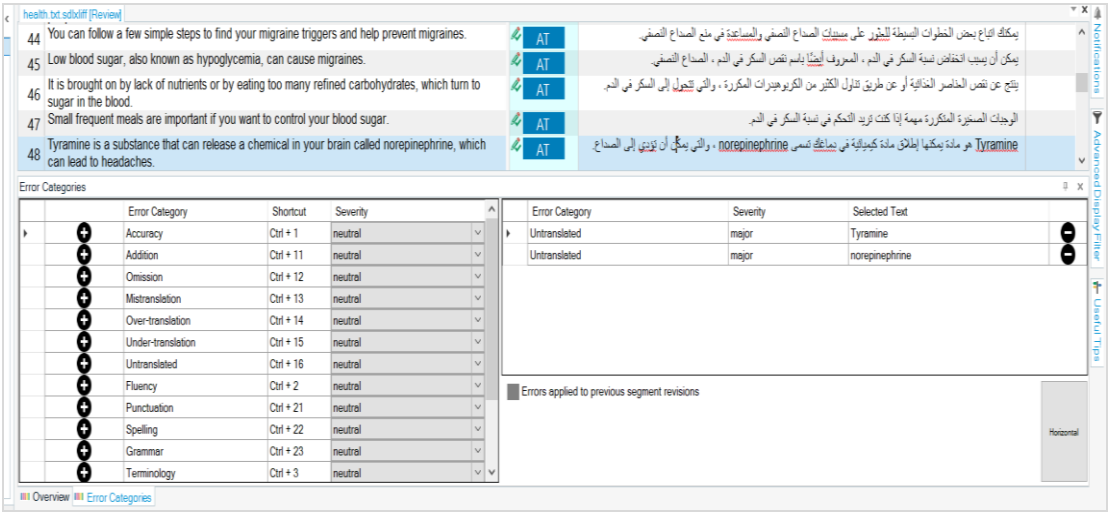
This study uses a corpus of six WikiHow articles randomly collected from three domains: cryptocurrency, cybersecurity, and healthcare. WikiHow articles represent the genre of instructional writing, which is characterized by the excessive use of imperative verbs and technical terms. The articles are user generated which means that they are not perfectly proofread, unlike the genres typically used to evaluate MT models. In total, the corpus consists of 6,641 word tokens and 2,116 word types.

5.2. Tools

Google Translate API, available through SDL Trados Studio 2019[3], is used because it offers both the statistical and neural models for English-Arabic translation and it is a "truly global product" with over 140 billion words translated every day and more than 1 billion monthly active users (Schuster, 2017). In addition, the DQF plugin for SDL Trados Studio 2019[4] is used. It is a software which connects projects created in Trados Studio to the TAUS Quality Dashboard. The taxonomy used in this plugin is based on the harmonized DQF-MQM error typology. Figure 1 below is a screenshot of the DQF annotation environment in SDL Trados 2019.

Figure 1

The DQF Annotation Environment in SDL Trados Studio 2019



5.3. Annotation

Errors in the output of each MT model are manually annotated segment-by-segment. For a deeper analysis, the errors are tagged on a subcategory level. Sometime, the same word exhibits more than one error type and each one is counted separately. The analysis is then exported to the Quality Dashboard where the annotations are download in Excel format. The analysis does not stop at the level of classifying errors. It delves deeper into each error category to figure out the patterns that trigger these errors.

6. Results

The results show that NMT surpasses the performance of SMT for the six articles under study, as illustrated in Figure 2 below. NMT achieves 79% error reduction compared to SMT (260 to 1227, respectively). It also

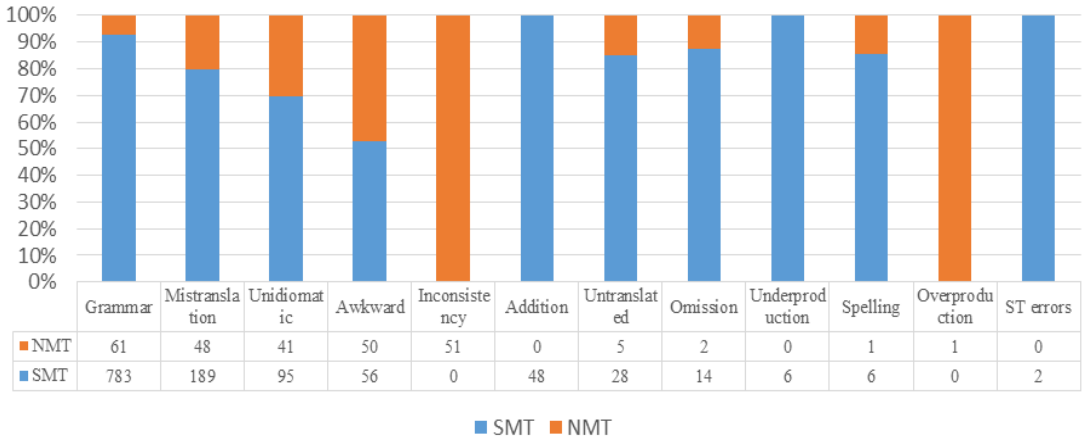
[3] Disclaimer: The machine-translated corpus under analysis was generated by Google Translate on March 15, 2020. Translation of the source text by the same MT engine might differ at the time of publication.

[4] https://appstore.sdl.com/language/app/taus-dqf-for-sdl-trados-studio/477/

produces fewer major errors than SMT; the number of major errors in the neural output is 85 errors against 575 in the statistical output. This means that the neural output is more accurate than the statistical one.

Figure 2

Number of Error Types in SMT and NMT Output



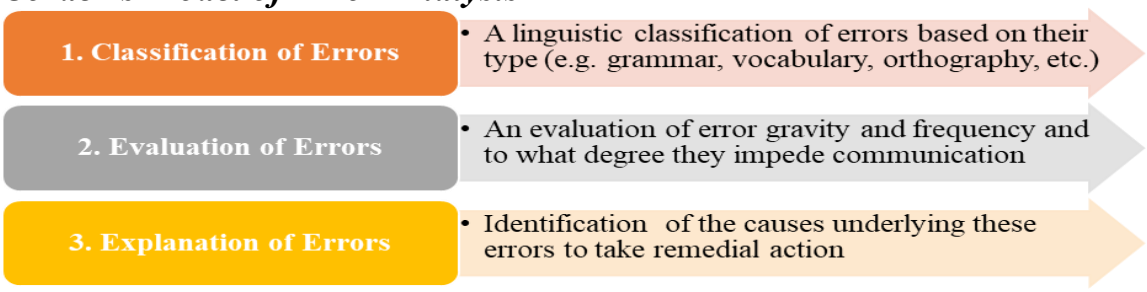
	Grammar	Mistranslation	Unidiomatic	Awkward	Inconsistency	Addition	Untranslated	Omission	Underproduction	Spelling	Overproduction	ST errors
NMT	61	48	41	50	51	0	5	2	0	1	1	0
SMT	783	189	95	56	0	48	28	14	6	6	0	2

6.1. Procedure

The procedure followed in the analysis is broadly guided by Corder's (1975) three-step model of error analysis in second-language acquisition as presented in Figure 3 below. However, some modifications were made to better reflect the different nature of the MT discipline. Section 6.2 below will discuss each error type, the patterns that cause such error, and whether using NMT alleviates the problem. Each finding is supported by an example in a table format. Each table contains a column for the source text, a second for the MT output and a third for the specific issue addressed. Examples may contain more than one error type, but only the one under discussion will be highlighted. The problematic source string is underlined and written in bold. Erroneous translation(s) are written in red font. When corrected, they are written in green font. Post-edited versions are sometimes provided when both models fail to provide a correct translation.

Figure 3

Corder's Model of Error Analysis



6.2. Error Analysis

6.2.1. Grammatical Errors

The error category NMT truly beats is grammatical mistakes. Using the neural model eliminates grammatical errors related to person agreement, case, dual form, exception, and conditionals. However, grammar remains the top error category in the output of both models.

The long-distance dependency caused by compounding and conjunction is the main culprit in many grammatical mistakes in the statistical output. It causes article mismatches, incorrect word order, and gender disagreement. This means that to yield better results from SMT, the source text better not include compounds or complex sentences. The neural model, however, handles long-distance dependencies better than the statistical one as example 1 shows.

Example 1

Source	SMT	Issue
Use a **strong, secure password** that would be difficult for anyone to guess.	استخدم **قوية وكلمة مرور آمنة** التي سيكون من الصعب على أي شخص تخمينها.	Incorrect word order
	NMT	
	استخدم **كلمة مرور قوية وآمنة** يصعب على أي شخص تخمينها.	
Caffeine is the prime ingredient in some **migraine relief medications**	SMT	Incorrect use of the definite article ال ‘al’
	الكافيين هو العنصر الرئيسي في **الأدوية تخفيف الصداع** بعض **النصفي**	
	NMT	
	الكافيين هو المكون الرئيسي في بعض **أدوية تخفيف الصداع النصفي**	
Unlike using **traditional finance systems**, whereby the system itself (like PayPal or a bank) is compensated with a fee, Bitcoin bypasses this entire system.	SMT	Gender disagreement
	خلافا باستخدام **نظم التمويل التقليدي**، حيث يتم تعويض النظام نفسه (مثل باي بال أو بنك) مع رسم، بيتكوين يتجاوز هذا النظام بأكمله.	
	NMT	
	على عكس استخدام **أنظمة التمويل التقليدية** ، حيث يتم تعويض النظام نفسه (مثل PayPal أو أحد البنوكBitcoin) برسوم ، فإن تتجاوز هذا النظام بأكمله.	

Incorrect parsing and part-of-speech tagging manifest in SMT. The model struggles with conjunctions. It tends to excessively insert the Arabic conjunction و incorrectly without any linguistic reason, even inserting two conjunctions in some cases. It also ignores the Arabic rules regarding the use of ف 'fa' before the apodosis[5] of a conditional (اقتران جواب الشرط بالفاء), for instance, when the apodosis is a verbal sentence denoting a command, request, prohibition, desire, or wish. SMT sometimes misinterprets conjunctions for other parts of speech such as mistaking the subordinate conjunctions "since" and "once" for adverbs. It also uses synonymous conjunctions without considering their different usage, for example, translating "because" always into "بسبب" even when it is followed by an independent clause in the translation. NMT, on the other hand, shows improved performance when translating conjunctions as example 2 presents.

Example 2

Source	SMT	Issue
The Bitcoin network is resistant to government regulation, **and** it has gained a loyal following among people who engage in illegal activities	شبكة بيتكوين المقاومة للتنظيم الحكومي، **وفقد** اكتسب التالية الموالية بين الناس الذين ينخرطون في أنشطة غير مشروعة NMT تقاوم التنظيم الحكومي Bitcoinإن شبكة ، **وقد** اكتسبت ولاءً مخلصًا بين الأشخاص الذين يشاركون في أنشطة غير مشروعة	Use of two conjunctions
If the scammer **is impersonating** a friend or family member rather than a business or government agency, **contact** that person directly.	SMT إذا المخادع **وينتحل** صديق أو أحد أفراد الأسرة بدلا من وكالة الأعمال أو الحكومة، **اتصل** هذا الشخص مباشرة. NMT إذا كان المحتال **ينتحل** شخصية صديق أو فرد من العائلة بدلاً من مؤسسة تجارية أو وكالة حكومية ، **فاتصل** بهذا الشخص مباشرةً.	- Incorrect addition of و - No ف in the apodosis
Since bright or flashing lights can sometimes lead to migraines, you should wear sunglasses on sunny days or even bright winter days.	SMT **منذ** مشرق أو الأضواء الساطعة يمكن أن يؤدي في بعض الأحيان إلى الصداع النصفي، يجب عليك ارتداء النظارات الشمسية في الأيام المشمسة أو أيام الشتاء حتى مشرق. NMT **نظرًا لأن** الأضواء الساطعة أو الوامضة يمكن أن تؤدي أحيانًا إلى الصداع النصفي ، يجب عليك ارتداء النظارات الشمسية في الأيام المشمسة أو حتى أيام الشتاء المشرقة.	Misinterpreting the part of speech for 'since'.

[5] The English references used for the translation of Arabic grammar terms are Wright (1996), Ryding (2005) and Sterling (2018).

Prepositions are most challenging for SMT. The model tends to translate their "meaning" rather than looking for the ones which collocate best with the translated neighboring words. This also leads the model to overlook inserting prepositions in the translation when the source does not include them. Preposition stranding is an additional difficulty for SMT. The model fails to insert the resumptive pronoun the Arabic structure requires, leaving the preposition dangling at the end of the Arabic sentence as it is in the English source. The neural model is much better at translating prepositions, even when they are stranded in English, as example 3 demonstrates. NMT generates 8 preposition errors against 95 for the statistical one.

Example 3

Source	SMT	Issue
Try to avoid all foods that you are allergic **to** as well as those you think you might be allergic **to**.	في محاولة لتجنب كل الأطعمة التي لديهم حساسية **لوكذلك** تلك التي تعتقد أنك قد تكون لديهم حساسية **ل**.	- Incorrect preposition collocates - Stranded preposition
	NMT	
	حاول أن تتجنب جميع الأطعمة التي تشكو من **حساسيتها** وكذلك تلك التي تعتقد أنك قد تكون مصابًا بالحساسية **منها**.	

The statistical output exhibits an extreme case of over-nominalization. Since the corpus of the analysis features the genre of instructional writing, imperative verbs are prevalent. SMT often translates the imperative as اسم (noun) or مصدر (verbal noun). This might be due to the fact that some human translators tend to translate the imperative mood using the imperative قم 'do' + the verbal noun of the imperative. Therefore, when SMT is trained with such data, it links the English imperative to the Arabic verbal noun; yet it neglects the insertion of the imperative قم. Moreover, verbs in the present simple are often incorrectly translated by SMT as verbal nouns. Again, the neural model is more efficient in this area as example 4 presents.

Example 4

Source	SMT	Issue
Copy the email address and **paste** it into a document	**نسخ** عنوان البريد الإلكتروني و**لصقه** في مستند	Nominalization of the imperative verbs
	NMT	
	انسخ عنوان البريد الإلكتروني و**الصقه** في مستند	
The code is good for a few minutes, then it **expires**.	SMT	Nominalization of the present simple verb
	رمز جيد لبضع دقائق، ثم **انتهاء** صلاحيته.	
	NMT	
	الرمز جيد لبضع دقائق ، ثم **تنتهي صلاحيته**.	

ISSN 1110-2721

Occasional Papers
Vol. 75: July (2021)

Although NMT yields better results than SMT in terms of gender agreement even with long-distance dependencies, it is still unable to determine the right gender of the pronoun 'you' from the local or global context, just like the statistical model as demonstrated in example 5. While the words "contraceptives" and "estrogen" are very indicative of the gender, both models translate the verbs in the masculine form, which means that they are incapable of sentence-level reasoning.

Example 5

Source	SMT	Issue
you might need to avoid or change the way you use oral contraceptives with estrogen	قد تحتاج إلى تجنب أو تغيير طريقة استخدام وسائل منع الحمل عن طريق الفم مع هرمون الاستروجين	Gender mismatch
	NMT	
	قد تحتاج إلى تجنب أو تغيير الطريقة التي تستخدم بها موانع الحمل الفموية التي تحتوي على هرمون الاستروجين	
	Post-edited Version	
	قد تحتاجين إلى تجنب أو تغيير الطريقة التي تستخدمين بها موانع الحمل الفموية التي تحتوي على هرمون الاستروجين	

6.2.2. Mistranslations

There are 189 mistranslations in the statistical output and 48 in the neural one. Although the number of errors is reduced, the percentage of mistranslations to the total number of errors is higher by 3% for NMT (18% compared to 15% for SMT).

The causes of mistranslation errors in SMT are almost the same in NMT; however, the neural model still shows better performance. Polysemy, idioms, technical terms, named entities and faulty training data are the top causes of mistranslations in the two models. But using the neural model eliminates mistranslations caused by phrasal verbs, compound adjectives, and ergative verbs in the statistical output as shown in example 6.

Example 6

Source	SMT	Issue
turn notifications **on** so that you'll be alerted when there's an update available.	تحويل الإخطارات على ذلك أن عليك أن تكون نبهت عندما يكون هناك تحديث متوفر	Mistranslation of the phrasal verb due to the long-distance dependency
	NMT	
	قم بـتشغيل الإشعارات حتى يتم تنبيهك عندما يتوفر تحديث	
Keeping your operating	SMT	

Source	SMT	Issue
system **up-to-date** ensures you have the strongest available security.	الحفاظ على نظام التشغيل الخاص بك ما يصل إلى تاريخ يضمن أن يكون لديك أقوى الأمان المتوفرة.	Mistranslation of the compound adjective
	NMT	
	يضمن تحديث نظام التشغيل الخاص بك أن يكون لديك أقوى أمان متاح.	
Even frozen fruits and vegetables can **benefit** your health.	SMT	Mistranslation of the ergative verb
	حتى الفواكه المجمدة والخضروات يمكن أن تستفيد صحتك.	
	NMT	
	حتى الفواكه والخضروات المجمدة يمكن أن تفيد صحتك.	

A feature that is unique to the neural model is the way it handles rare or out-of-vocabulary words. One of the many quirks of NMT is that it always tries to figure out the meaning of the word even if it has not seen this word before in the training data. In example 7, both "feverfew" and "butterbur" are mistranslated as حمى (fever) and زبدة (butter). This indicates that NMT uses sub-word sequences to overcome the rare word challenge. It splits the first word into "fever"+ "few" and the second into "butter" + "bur". Still, this does not explain why it translates "fever-" and "butter-" but not "-few" and "-bur", which are also valid words. This implies that NMT might be using the expectation maximization algorithm (Koehn, 2020, p. 228) which prefers longer sub-words over shorter ones and removes the sub-word with the least probability.

Example 7

Source	NMT	Issue
Extracts of the **feverfew** and **butterbur** plants and kudzu root could possibly help.	يمكن أن تساعد مقتطفات نباتات الحمى والزبدة وجذر كودزو.	Mistranslation due to sub-wording
	Post-edited Version	
	يمكن لخلاصة نباتات الأقحوان والأرام وجذر الكودزو التخفيف من الصداع النصفي.	

The neural model also tends to sacrifice accuracy to achieve a fluent output when faced with rare words (Koehn and Knowles, 2017). It sometimes produces neologisms only to preserve the structure of the output sentence. In example 8, NMT translates "pounding" as خفقاني, a word that does not exist in the Arabic lexicon. In Arabic, one way to derive adjectives is by attaching ـيّ suffix to the masculine noun or ـيّة to the feminine noun, for instance, مصر (Egypt) and مصريّ (Egyptian). This

suffix is known as ياء النسب (relational yāʾ) and functions in a manner similar to the -ian suffix in English. In a futile attempt to produce an equivalent target adjective, NMT mimics this rule by adding the suffix to the noun خفقان.

Example 8

Source	NMT	Issue
The pain is described as a **pounding**, pulsating, throbbing headache.	يوصف الألم بأنه صداع **خفقاني** ، نابض ، خفقان. Post-edited Version يوصف الألم المصاحب لهذا الصداع بأنه يشبه **الخبط** أو النبض أو الخفقان.	Neologism

6.2.3. Unidiomatic Style

The unidiomatic style means that the translation is comprehensible but unnatural. This category exclusively covers collocation errors in the two outputs. To ensure that the error annotation in this category is not guided merely by personal stylistic preferences, *Dar El-Ilm's Dictionary of Collocations* (Ghazala, 2007), المدونة اللغوية العربية لمدينة الملك عبدالعزيز للعلوم والتقنية (KACST Arabic Corpus)[6], and Google counts are used.

The neural model is better than the statistical one at selecting collocations, reducing errors from 95 to 41. Both models, however, tend to translate the "meaning" rather than looking for the adequate Arabic collocation. Almost 50% of the errors in this category are noun collocations, followed by verb collocations and finally adjective collocations. Example 9 presents some cases in which NMT yields better collocations in Arabic.

Example 9

Source	SMT	Issue
Some exchanges allow you **to make a deposit** in person to their bank account.	بعض التبادلات تسمح لك **لجعل وديعة** في شخص إلى حساباتهم المصرفية. NMT تسمح لك بعض البورصات **بعمل إيداع** شخصي في حسابها المصرفي.	Unidiomatic noun collocation
If you **have established** which foods seem to trigger your migraines, **eliminate** them from your diet.	SMT إذا كنت قد **أنشأت** الأطعمة التي يبدو أنها تثير الصداع النصفي، و**القضاء عليها** من النظام الغذائي الخاص بك. NMT إذا كنت قد **حددت** الأطعمة التي يبدو أنها تسبب الصداع النصفي ، **فقم بإزالتها** من نظامك الغذائي.	- Incorrect part-of-speech - Unidiomatic verb collocations

[6] https://corpus.kacst.edu.sa/collocation.jsp

Source	SMT	Issue
The company currently has only web platform which is also **mobile friendly**.	SMT وتمتلك الشركة حاليا على شبكة الإنترنت منصة الوحيدة التي هي أيضا صديقة للجوال. NMT لدى الشركة حاليًا منصة ويب فقط وهي أيضًا مناسبة للجوّال.	Unidiomatic adjective collocation

6.2.4. Awkward Style

This category describes errors that do not hinder the meaning but make the output wordy and sometimes difficult to follow. Using the neural model does not improve the translation quality in this area; both models produce almost the same number of errors.

Verbose training data is the main cause of generating wordy translations in both models. Patterns are observed in the translation of the possessive pronoun "your", the passive voice, and the imperative. Driven by a misconceived notion of faithfulness to the source text, human translators sometimes mimic the English structure and use expressions such as الخاص أو الخاصة بك to translate the possessive pronoun and ignore that Arabic uses the attached possessive pronoun ـك. They also tend to translate "by" in the passive structure as من قبل وبواسطة instead of reverting the voice back into active. In addition, they prefer to translate the passive verb into تم+the verbal noun instead of using the Arabic passive form of the verb.

Moreover, humans sometimes translate imperative verbs into the قم + the verbal noun of the imperative. However, using the imperative form in Arabic is more natural and economical. But since the Arabic imperative makes heavy use of diacritics, some translators find it difficult and prefer being on the safe side by using the verbal noun. The structure has also been widely used as a band-aid solution to the problem of Arabic diacritics which have not been largely supported in desktop publishing.

Source-text interference is another cause of verbose language. Like human translators, both models try to remain as close to the source text as possible. This sometimes results into insertion of unnecessary words. In example 10, for instance, the noun "paper" is uncountable, so the word "piece" is used to express singularity. In Arabic, however, there are singular, dual, and plural forms of the word "paper". It is unnecessary to translate the word "piece". Therefore, instead of the three-worded phraseقطعة من الورق, it is more economical to use a single word ورقة.

Example 10

Source	SMT	Issue
The image is printed on a long **piece of paper**	تتم طباعة الصورة على قطعة طويلة من الورق	Verbose language
	NMT	
	تتم طباعة الصورة على قطعة طويلة من الورق	
	Post-edited Version	
	تُطبَع الصورة على ورقة طويلة	

Synonymy is the third cause of wordy translations in the two outputs. It causes both models to repeat the same translation. This is particularly observed in medical term of Latin origin as example 11 illustrates. In cases like this, it is better to either transliterate the technical term or omit it from the translation.

Example 11

Source	SMT	Issue
Low blood sugar, also known as **hypoglycemia**, can cause migraines.	انخفاض نسبة السكر في الدم، والمعروف أيضا باسم نقص السكر في الدم، يمكن أن يسبب الصداع النصفي.	Repetition of the same translation
	NMT	
	يمكن أن يسبب انخفاض نسبة السكر في الدم ، المعروف أيضًا باسم نقص السكر في الدم ، الصداع النصفي.	
	Post-edited Version	
	يمكن لانخفاض السكر في الدم والمعروف أيضًا بالهايبوجلايسيميا أن يسبب الصداع النصفي.	

6.2.5. Inconsistency

Exclusive to the neural output, 51 errors are tagged under this category, all in the cryptocurrency domain. Inconsistency means that the same term is translated differently throughout the text.

The statistical model is nothing but consistent, even in the type of errors it produces. However, the neural model shows inconsistency when translating the same term. For instance, it translates the plural word "Bitcoins" in three different ways: البيتكوين, عملات البيتكوين, and the untranslated "bitcoins", as example 12 shows. In some cases, it translates the abbreviation "ATMs" as أجهزة الصراف الآلي while leaving it untranslated in others. This means the neural model has different word embeddings for the same word, which results into this inconsistency.

Example 12

Source	NMT	Issue
One of **Bitcoins** popular uses is as an investment	أحد الاستخدامات الشائعة لـ Bitcoins هو استثمار	Inconsistent translation of the same word
Always back up your wallet to an external hard drive to avoid losing your **Bitcoins**.	NMT قم دائمًا بعمل نسخة احتياطية من محفظتك على محرك أقراص صلبة خارجي لتجنب فقدان عملات البيتكوين الخاصة بك.	
Access the codes needed from your account via your smartphone to load **bitcoins** onto your wallet.	NMT قم بالوصول إلى الرموز المطلوبة من حسابك عبر هاتفك الذكي لتحميل البيتكوين على محفظتك.	

6.2.6. Addition Errors

There are 48 addition errors in the statistical output. However, using the neural model eliminated this category completely. SMT inserts chunks from the Holy Quran and reporting verbs as presented in example 13. This can be attributed to misalignment in the phrase table of the statistical model.

Example 13

Source	SMT	Issue
When you meet the seller face-to-face, you will need to access your Bitcoin wallet via your smartphone, tablet, or laptop.	فإذا لقيتم الذين البائع وجها لوجه، وسوف تحتاج إلى الوصول بيتكوين محفظتك عبر هاتفك الذكي أو الجهاز اللوحي، أو الكمبيوتر المحمول.	Addition of Quranic text
	NMT عندما تقابل البائع وجهاً لوجه ، ستحتاج إلى الوصول إلى محفظة Bitcoin عبر هاتفك الذكي أو جهازك اللوحي أو الكمبيوتر المحمول.	
When you visit your doctor, he will check to see whether your skin is abnormally pale.	SMT عند زيارة الطبيب، وقال انه سوف تحقق لمعرفة ما إذا كانت بشرتك غير طبيعي شاحب.	Addition of a reporting verb
	NMT عندما تزور طبيبك ، سيتحقق لمعرفة ما إذا كانت بشرتك شاحبة بشكل غير طبيعي.	

Sometimes, addition errors happen because a certain structure is more probable than the other. In example 14, SMT translated the verb phrase

"do eat" into the negative form in Arabic لا تأكل (do not eat) by adding the negation particle لا.

Example 14

Source	SMT	Issue
If you **do eat** meats, make sure any beef is lean	إذا كنت لا تأكل اللحوم والتأكد من أي لحوم البقر الخالية من الدهون	Addition of the negation particle
	NMT	
	إذا كنت تأكل اللحوم ، فتأكد من أن أي لحم بقري قليل الدهن	

6.2.7. Untranslated Words

This category reports issues where English words are found in the Arabic output. It does not include instances of Do-Not-Translate items such as brand names, where it is acceptable to leave the word untranslated.

SMT produces more untranslated words than NMT (28 to 25). Unlike NMT, when the statistical model faces a word that does not exist in its training data, it spits it back untranslated as example 15 demonstrates.

Example 15

Source	SMT	Issue
Extracts of the feverfew and **butterbur** plants and kudzu root could possibly help.	مقتطفات من الينسون والنباتات butterbur وجذر كودزو يمكن ربما المساعدة.	Untranslated word
	Post-edited Version	
	يمكن لخلاصة نباتات الأقحوان والأرام وجذر كودزو التخفيف من الصداع النصفي.	

True-casing is also the cause of some untranslated words, especially in the statistical output. True-casing is the task of inferring the correct case of a word to distinguish named entities from regular nouns. All CAPs words sometimes trick SMT into treating them as Do-Not-Translate items as demonstrated in example 16.

Example 16

Source	SMT	Issue
Bitcoin usage does not require a name, or any other personal information, simply an **ID** for your digital wallet	بيتكوين الاستخدام لا يحتاج الى اسم، أو أي معلومات شخصية للمحفظة IDأخرى، مجرد الرقمية الخاصة بك	Untranslated word
	Post-edited Version	
	لا يتطلب استخدام البيتكوين اسمًا أو أي معلومات شخصية أخرى سوى الهوية المرتبطة بمحفظتك الرقمية	

6.2.8. Omission Errors

Using the neural model almost eliminates omission errors, generating only two errors compared to 14 in the statistical output. These are piece of information that are present in the source text but missing in the target.

Misalignments are probably the number one cause of omission errors in the statistical output. The most probable cause of such deletions is that they are aligned to the NULL value. A NULL is an empty category element in the model's parse tree that does not have a correspondent in the target text . After the alignment process, any unaligned word or phrase is mapped to NULL, therefore, dropped from the translation. In example 17, the verb "forward" is omitted from the output.

Example 17

Source	SMT	Issue
In the US, you can also **forward** to spam@uce.gov.	في الولايات المتحدة، يمكنك أيضا إلىspam@uce.gov.	Omitted verb
	NMT	
	في الولايات المتحدة ، يمكنك أيضًا إعادة التوجيه إلى spam@uce.gov.	

As for NMT, the attention model used might be the cause of the two instances of omission in the neural output. Attention in NMT plays the role of alignment in SMT. To achieve more fluency for long sentences, NMT system uses attention models, which focus the view of the input sentence on what the system deems as *important* words (See Koehn, 2020). However, this sometimes causes omissions in the target text. In example 18, GNMT omitted the adverb "poorly" from the output, causing false repetition and loss of meaning.

Example 18

Source	NMT	Issue
Untreated or **poorly** treated hyperthyroidism can lead to heart problems	يمكن أن يؤدي فرط نشاط الغدة الدرقية غير المعالج أو غير المعالج إلى مشاكل في القلب	Omitted adverb
	Post-edited Version	
	يمكن أن يؤدي فرط نشاط الغدة الدرقية غير المعالج أو المعالج بصورة خاطئة إلى مشاكل في القلب	

6.2.9. Miscellaneous

Other less frequent errors are observed in both models. There are six instances of underproduction where SMT inserts incomplete target terms. There are six spelling mistakes in the statistical output, compared to only

one in the neural output. The neural model is more capable of overcoming source-texts typos whereas the statistical one produces erroneous translations as a result. There is only one case of overproduction and it is generated by NMT. Overproduction means that the same target term is repeated. This is not to be confused with addition errors where a target term that does not exist in the source is inserted in the output.

7. Conclusion, Final Remarks and Future Research

In this study, a comparison is made between the quality of SMT and that of NMT based on a detailed error analysis. Results show that NMT outperforms SMT for the English-into-Arabic translation across all tested domains. The neural model has indeed reduced the number of errors generated in the Arabic output by almost 80%; and this answers research question no. 1 (RQ1). In response to RQ2, NMT has produced less severe errors than SMT. The number of major errors in the neural output is 85 errors against 575 in the statistical output. As for the error types addressed in RQ3, both models have produced grammatical errors, mistranslations, unidiomatic and verbose language, untranslated words, omissions, and spelling mistakes. However, it is important to note that NMT has considerably reduced the number of these error types in the Arabic output. Regarding RQ4, the neural model has eliminated some error types reported in the statistical output. These errors are additions, underproduction, and mistranslations caused by source text errors. NMT has also overcome errors related to person agreement, case, dual forms, exceptions, conditionals, phrasal verbs, and ergative verbs. The neural model has reported cases of inconsistency and overproduction which did not exist in the statistical output, which responds to RQ5.

This paragraph sums up the patterns that trigger each error type in response to RQ6. The main causes of **grammatical errors** in both models, especially the statistical one, are compounds, function words, imperatives, present simple verbs, and the second person pronoun “you” as examples 1-5 present. The **mistranslations** in both outputs are caused by polysemous words, idioms, technical terms, named entities, and faulty training data. Out-of-vocabulary words also cause mistranslations in the neural output as shown in example 7. Collocations, especially noun collocations, are the primary reason for the **unidiomatic language** in both models as evident in example 9. **Verbose language** in both outputs is observed in the translation of possessive pronouns, the passive voice, imperatives, and quantifiers used with uncountable nouns. Using synonyms in the same sentence also leads both models to repeat the same translation twice. The **inconsistencies** found in the neural output, as highlighted in example 12, can be attributed to the word having different

embeddings in the model's corpus. **Additions** in the statistical output are mainly caused by misalignment of source and target phrases in the model's training data and the emphatic use of auxiliaries in affirmative sentences, as noted in examples 13 and 14. Out-of-vocabulary words and incorrect true-casing are the top causes of **untranslated words** in the Arabic outputs, especially the statistical one as shown in examples 15 and 16. **Omissions** are caused by misalignment in SMT and attention in NMT.

In conclusion, NMT definitely shows more promise than SMT in terms of quality. It produces less errors and more fluent output than SMT. However, NMT does not seem to be perfect, and is not expected to replace human translators anytime soon. It can bring considerable gains in terms of productivity and cost when combined with post-editing. It is also worth mentioning that the "improved" quality of the neural output sometimes means that errors are harder to spot; one can be easily swayed by its fluency to the extent of letting errors pass uncritically. Does this mean more post-editing time for the neural output? Does it mean a higher skill set for post-editors? These are questions the study intends to answer in future research.

References

Abdelaal, N. M., & Alazzawie, A. (2020). Machine Translation: The Case of Arabic-English Translation of News Texts. *Theory and Practice in Language Studies*, 10(4), 408-418.

Almahairi, A., Cho, K., Habash, N., & Courville, A. (2016). First result on Arabic neural machine translation. *arXiv preprint arXiv:1606.02680.*

Alrajeh, A. (2018). A Recipe for Arabic-English Neural Machine Translation. *arXiv preprint arXiv:1808.06116.*

Cadwell, P., O'Brien, S., & Teixeira, C. S. (2018). Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 301-321. doi:10.1080/0907676X.2017.1337210

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In D. McCarthy, & S. Wintner (Ed.), *11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 249-256). Trento: Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 136-158). Prague: Association for Computational Linguistics.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Waya, A. (2017). Is Neural Machine Translation the New State of the Art? *The Prague Bulletin of Mathematical Linguistics*, (108), 109-120.

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (Vol. I, pp. 9-38). Springer International Publishing. Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-319-91241-7.pdf

Charniak, E., Knight, K., & Yamada, K. (2003). Syntax-based Language Models for Statistical Machine Translation. *Proceedings of the Ninth Machine Translation Summit*, (pp. 40-46). New Orleans.

Corder, S. (1975). Error Analysis, Interlanguage and Second Language Acquisition. *Language Teaching & Linguistics*, *8*(4), (pp. 201-218). doi:10.1017/S0261444800002822

Culy, C., & Riehemann., S. Z. (2003). The Limits of N-Gram Translation Evaluation Metrics. *Proceedings of the Ninth Machine Translation Summit*, (pp. 71-78). New Orleans.

Dreyer, M., Hall, K., & Khudanpur, S. (2007). Comparing reordering constraints for SMT using efficient BLEU oracle computation.

In *Proceedings of SSST, NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation* (pp. 103-110).

Durrani, N., Dalvi, F., Sajjad, H., & Vogel, S. (2017). QCRI machine translation systems for IWSLT 16. *arXiv preprint arXiv:1701.03924.*

Esperança-Rodier, E., Brunet-Manquat, F., & Eady, S. (2019, November). ACCOLÉ: A Collaborative Platform of Error Annotation for Aligned Corpora. In *Translating and the computer 41*. London: AsLing. Retrieved from https://www.asling.org/tc41/wp-content/uploads/TC41-Proceedings.pdf

Ghazala, H. (2007). *Dar El-Ilm's Dictionary of Collocations*. Beirut: Dar El Ilm Lilmalayin.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L.,… Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv preprint arXiv:1803.05567.*

Hossain, M. M., Anastasopoulos, A., Blanco, E., & Palmer, A. (2020). It's not a Non-Issue: Negation as a Source of Error in Machine Translation. *arXiv preprint arXiv:2010.05432.*

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. *arXiv preprint arXiv:1610.01108.*

Klubička, F., Toral Ruiz, A., & Sánchez-Cartagena, M. V. (2017). Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics* No. 108 (pp. 121-132).

Koehn, P., & Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation* (pp. 102-121). New York City: Association for Computational Linguistics.

Koehn, P., & Knowles, R. (2017, August). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28-39). Vancouver: Association for Computational Linguistics.

Koehn, P. (2020). *Neural machine translation*. Cambridge: Cambridge University Press.

Linares, J. A. (2008). Empirical machine translation and its evaluation [Doctoral thesis, Universitat Politècnica de Catalunya]. Citeseerx. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.6658&rep=rep1&type=pdf

Lommel, A. (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In J. Moorkens, S. Castilho, F. Gaspari,

& S. Doherty (Eds.), *Translation Quality Assessment: From Principles to Practice* (Vol. I, pp. 109-127). Springer International Publishing.

Lommel, A. (2020, July 22). Arabic: To Localize or Not to Localize, That Is the Question. *CSA Research*. Retrieved from https://csa-research.com/Blogs-Events/Blog/localizing-translating-arabic-websites

Marie, B., Fujita, A., & Rubino, R. (2021). Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 7297–7306). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.acl-long.566.pdf

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Popović, M. (2017). Comparing Language Related Issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, (108), 209-220.

Prospering in the pandemic: the top 100 companies. (2020, June 19). *Financial Times*. Retrieved from https://www.ft.com/content/844ed28c-8074-4856-bde0-20f3bf4cd8f0

Ryding, K. C. (2005). *A Reference Grammar of Modern Standard Arabic*. New York: Cambridge University Press.

Sampson, G. (2003). *International Encyclopedia of Linguistics*. New York: Oxford University Press.

Schuster, M. (2017). Moving to Neural Machine Translation at Google [PowerPoint]. Retrieved from http://asru2017.org/Slides/AS17_InvitedTalk_Schuster.pdf.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007, June). Rule-based translation with statistical phrase-based post-editing. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 203-206).

Stasimioti, M., & Sosoni, V. (2020). MT output and post-editing effort: Insights from a comparative analysis of SMT and NMT output and implications for training. *Fit-For-Market Translator and Interpreter Training in a Digital Age*, 151.

Sterling, R. (2018). *A Grammar of the Arabic Language*. Oxon: Routledge.

Stroppa, N., Owczarzak, K., & Way, A. (2007). A Cluster-Based Representation for Multi-System MT Evaluation. *Proceedings of The 11th Conference on Theoretical and Methodological Issues in Machine Translation*, (pp. 114-121). Skövde.

TAUS. (n.d.). *Harmonized DQF-MQM Error Typology*. Retrieved from https://www.taus.net/qt21-project#harmonized-error-typology

Toral, A., & Sánchez-Cartagena, V. M. (2017, April). A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 1063-1073).

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. *arXiv preprint arXiv:1808.10432*.

Wright, W. (Trans.). (1996). *A Grammar of the Arabic Language* (3rd ed., Vol. II). Beirut: Librairie du Liban.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H.,... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*

List of Abbreviations

Abbreviation	Full Term
BLEU	Bilingual Evaluation Understudy
DFKI	German Research Center for Artificial Intelligence
DQF	Dynamic Quality Framework
LSP	Language Service Provider
MQM	Multidimensional Quality Metrics
MT	Machine Translation
NMT	Neural Machine Translation
RQ	Research Question
SMT	Statistical Machine Translation
TAUS	Translation Automation User Society
WER	Word Error Rate