Human Evaluation of Neural Machine Translation: The Case of *Deep Learning*

Marie Escribe
Guildhall School of Business and Law
London Metropolitan University
escribe.marie@gmail.com

Abstract

Recent advances in artificial neural networks now have a great impact on translation technology. A considerable achievement was reached in this field with the publication of *L'Apprentissage Profond*. This book, originally written in English (*Deep Learning*), was entirely machine-translated into French and post-edited by several experts. In this context, it appears essential to have a clear vision of the performance of MT tools. Providing an evaluation of NMT is precisely the aim of the present research paper. To accomplish this objective, a framework for error categorisation was built and a comparative analysis of the raw translation output and the post-edited version was performed with the purpose of identifying recurring patterns of errors. The findings showed that even though some grammatical errors were spotted, the output was generally correct from a linguistic point of view. The most recurring errors are linked to the specialised terminology employed in this book. Further errors include parts of text that were not translated as well as edits based on stylistic preferences. The major part of the output was not acceptable as such and required several edits per segment, but some sentences were of publishable quality and were therefore left untouched in the final version.

1 Introduction

The concept of a computer system designed for translation assistance is several decades old. While the first computers were created just before World War II to perform calculations (in ballistics), it quickly became apparent that they could be used as decoding tools (to decipher enemy encrypted transmissions such as the Enigma code). This achievement is often considered as one of the first steps towards Machine Translation (MT) (Planas, 2017). Obviously, translation is not exactly a matter of deciphering codes, but rather raises issues of equivalence between languages. However, this paved the way for MT, and experts began to build more and more tools (Rule-Based MT, Statistical MT). While the first studies on the use of neural networks for MT dates back to the 1990s (Ñeco and Forcada, 1997), Neural Machine Translation (NMT) has largely benefited from the advances in Artificial Intelligence (AI) and has thus grown considerably in recent years.

In November 2018, the first book translated by a NMT system (English>French), *L'Apprentissage Profond*, was published in France. The title says it all: *Deep Learning* is the very promising technology based on artificial neural networks used by Quantmetry and DeepL GmbH to translate this book. This has been widely publicised in both national and international media and it is often referred to as the first book entirely translated by an AI system (Zaffagni, 2018), since the amount of post-editing prior to the publication of this book is considered to be minimal. These advances in MT technology sometimes lead professionals to think that their jobs will entirely be performed by machines in the coming years: Bawa-Mason et al. (2018) pointed out that 38% of practising translators are worried that MT tools will end up replacing them. In this context, it appears crucial to conduct research in the area of recent MT systems in order to have a clear vision of the performance of NMT nowadays. This is precisely the objective of the present study. Implementing quality assessment methods is essential to monitor the evolution of MT systems. This is why several quality assessment frameworks have been

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 36–46.
https://doi.org/10.26615/issn.2683-0078.2019_005

proposed, including both human judgment and automatic metrics.

This research project proposes a method for the analysis of NMT output based on human evaluation. The aim is to establish a comparative study between the raw translation output and the post-edited version of *Deep Learning* in order to identify and analyse differences between the two versions. The edits performed were thus quantified and classified in order to identify recurring patterns of errors. The analysis of the outcomes obtained allowed to determine typical situations in which the performance of NMT is still insufficient.

2 MT Evaluation

2.1 Automatic Metrics vs Human Judgment

With the increasing development of MT systems, it became necessary to implement assessment techniques to evaluate the translations obtained and thus design more efficient systems. As a matter of fact, MT evaluation became a field in its own right.

Many scholars claim that automatic metrics are the most efficient solution because they are objective, fast and inexpensive compared to human evaluation. Among the many automatic metrics created, BLEU (BiLingual Evaluation Understudy) appears to be the most popular (Do, 2011) because it is considered to provide very accurate results that are strongly correlated with human judgments (Papineni et al., 2002). Similar metrics include the NIST metric (National Institute of Standards and Technology, Doddington, 2002) and METEOR (Metric for Evaluation of Translation with Explicit Ordering, Banerjee and Lavie, 2005). Other metrics are based on the error rate and the Levenshtein distance, such as the WER (Word Error Rate) score and the improved versions of this metric – i.e. PER (Position-independent Word Error Rate, Tillmann et al., 1997), TER (Translation Edit Rate or Translation Error Rate, Snover et al., 2006) and HTER (Human-targeted Translation Edit Rate, Snover et al., 2006).

The popularity of such metrics can be explained by the weaknesses of human evaluation. Having human evaluators judge a MT output, either by rating it or by post-editing it according to a reference, is a difficult task because such techniques are time-consuming, rather expensive and generally not re-usable. Moreover, such studies are highly subjective, as human evaluators do not necessarily agree on the quality of the MT output. In addition to this, error categorisation is a particularly difficult task when it comes to human evaluation.

Despite this, human judgment is paramount for designing effective evaluation systems and interpreting the scores they provide. The human input is crucial when it comes to informing experts in order to improve MT evaluation systems, since human analyses often serve as a framework for the creation of such tools. Vilar et al. (2006) argued that the interpretation of scores provided by automatic metrics can sometimes be unclear and that error classification and analysis by humans is therefore needed. Turian et al. (2003) also insisted on the importance of human judgment. In fact, several experts disagreed with the statement that automatic metrics show a good correlation with human judgments (Doddington, 2002; Callison-Burch et al., 2006). In this regard, Sennrich (cited in Pan, 2016) also pointed out that BLEU only focuses on precision and does not consider syntactic structures and grammar. Furthermore, Tinsley (cited in Pan, 2016), noted that BLEU scores are not efficient when it comes to evaluation of NMT.

The limitations of automatic metrics therefore make human judgment extremely valuable. Only human evaluators can tell whether the type of language used is adequate according to the context (register) or if a change in grammar or lexis at the post-editing stage is considerably affecting the meaning of a sentence. Indeed, Ulitkin (2013), who tested several automatic metrics such as BLEU and TER, stated that these tools could not provide quality assessment at the semantic or pragmatic levels. Consequently, it is necessary to conduct human evaluation of NMT output. Such methods usually focus on adequacy (i.e. whether the meaning has been rendered correctly) and fluency (i.e. grammaticality and fluency of the output) (Lavie, 2011) and generally require to elaborate an error classification.

2.2 Previous Work

Llitjós et al. (2005), who aimed to find an automation process for post-editing, were among the first experts to present an error typology. The classification they proposed served as a model for

that presented by Vilar et al. (2006) for human evaluation of SMT. These two classifications are indeed very similar, with three categories in common ("missing word", "word order" and "incorrect words"). Vilar et al. (2006) used a more comprehensive typology, with more sub-categories, thus allowing a more precise error identification. For instance, the sub-category "sense" (belonging to "word order") has, in turn, been divided into two categories (namely "wrong lexical choice" and "incorrect disambiguation"). Daems et al. (2017) also came up with a typology including similar categories – even though the aim of their study was to quantify the post-editing effort. Although the general classification appears to be different, these frameworks share common features (for instance the "lexicon" category in Daems et al., 2017 is similar to that of "wrong lexical choice" in Vilar et al., 2006).

It is important to note that these typologies were established before the creation of NMT, and it could therefore be argued that they concentrate mostly on features for which more recent MT systems are not likely to produce errors (even though their study was conducted in 2017, Daems et al. worked with a SMT system). Isabelle et al. (2017) argued that the performance of NMT was outstanding compared to other MT systems and, for this reason, one can think that the classifications mentioned above are now antiquated and cannot be used for NMT evaluation. However, this is not the case. Ahrenberg (2017), who established a comparison of a NMT output and a human translation, mostly built on Vilar et al. (2006) to create an error typology and even acknowledged that five categories out of six are directly inspired by their taxonomy. The framework for human analysis of NMT used by Hassan et al. (2018) also shares several features with those introduced before, with categories such as "missing word" and "word order" that were already present in the study of Vilar et al. (2006). These error types can also be found in Popovic (2018). Moreover, these typologies share a number of categories with several guidelines for post-editing. For instance, DePalma (2013) presented a categorisation (adapted from LISA QA Model) explaining the differences between "light" and "full" post-editing. Some of the errors that should be addressed by post-editors are similar to the categories mentioned above (with "omissions"/"additions" corresponding to "missing word"/"extra word" in Llitjós et al., 2005, for example). Further studies, such as the 'QT21 Harmonised Metric' (Lommel et al., 2015) and 'From Quality Evaluation to Business Intelligence' (TAUS Quality Dashboard, 2016) introduced DQF (Dynamic Quality Framework) tools allowing users to categorise and count errors segment-by-segment using issue type hierarchies (i.e. error typologies). Here again, several error categories are identical to other frameworks mentioned before (for instance, "addition", "omission", "punctuation", "spelling" and "grammar").

2.3 Performance of NMT

Isabelle et al. (2017) tested NMT systems with particularly challenging linguistic material and pointed out the cases in which NMT failed to provide a satisfying output thanks to a specific error typology. This study proved the efficiency of NMT over other MT systems and provided a list of strengths (such as the capacity to overcome many limitations of *n*-gram language modelling) and weaknesses (such as the translation of idioms) of NMT. However, some experts, such as Hassan et al. (2018), argue that the performance of NMT now equals human quality. This study proved to be highly controversial as other experts criticised their approach, especially regarding the definition of human parity (as pointed out by Diño, 2018). The authors claimed that human parity is achieved if the output is considered to be equivalent to a human translation according to bilingual human judges. This definition can appear as not rigorous enough, in particular when compared to other metrics (such as BLEU) in which human parity is achieved only if a candidate translation is completely identical to a translation produced by a human.

Of course, this discrepancy is due to an intrinsic problem in Translation Studies. The concept of equivalence itself is a controversial topic in the field (Hatim and Munday, 2004), and, very often, there is not only one possible translation for a given sentence, but rather several valid options. Therefore, establishing a comparison based only on a limited number of possible translations seems restrictive. On the other hand, evaluations established by human judges allow for more possibilities to be included, but they are subjective.

Several studies brought nuances to the findings of Hassan et al. (2018). Amongst them, Läubli et al. (2018) suggest that given the good quality of NMT output at the sentence level, analyses of NMT should focus on the document level. This suggestion was also made by Toral et al. (2018), who argue that important variables were not considered in the experiment of Hassan et al. (such as the languages involved, the translation expertise of the evaluators, etc.).

In fact, the need for MT evaluation is more important than ever with the development of NMT systems. They have become more and more popular in recent years, in particular because they are able to produce translations of high quality, compared to Statistical MT, as pointed out by Sennrich (2016). Furthermore, NMT is a relatively recent technology, whereas most automatic evaluation metrics were created more than 15 years ago. Consequently, it appears relevant to conduct human evaluation of NMT output in order to identify recurrent error patterns and thus to investigate how to integrate the recognition of such patterns in automatic metrics. Ahrenberg (2017) stressed the fact that Translation Studies and MT evaluation have mostly evolved separately and therefore lack common terminology. This is unfortunate because cooperation between translators and computer engineers is paramount to create efficient evaluation systems, since knowledge from linguists is important feedback for the creation of adequate assessment methods. This is particularly true when it comes to NMT. For instance, Monti et al. (2018, pp. 19-20) pointed out that only a few studies were implemented regarding multiword units NMT output, and further research is therefore needed.

3 Experimental Setup

3.1 Material Investigated

The material investigated for this project consists of an excerpt from *Deep Learning* (Goodfellow, Bengio and Courville, 2016). This manual is extremely comprehensive, which is why it is known to be a 'must-have' for Data Science students or practitioners aiming to use deep learning models and is recommended by several universities, even in non-English-speaking countries (Bousquet, 2018). However, the English language can be perceived as a barrier to a full understanding of the book. On the other hand, translating this monumental work (800 pages) would be both long and expensive – estimations showed that it would require approximately an entire year of work and up to 150,000 euros (Stora, cited in Zaffagni, 2018). Quantmetry and DeepL GmbH came up with a bold solution to this problem – translating *Deep Learning* by using deep learning methods. This incredible *mise en abîme* was successful, as *L'Apprentissage Profond*, the French translation of *Deep Learning*, was published in 2018 and this achievement received strong media attention. To do this, the developers had to create a glossary of 200 specialised terms (Zaffagni, 2018) and to implement a tool capable of handling LaTeX format. The system thus developed showed impressive results, as the book was translated in no more than 12 hours, for a total budget of 30,000 euros, including printing (Bousquet, 2018). The translation was then entirely post-edited by several experts from the ENSAI, INRIA and CNRS [1] (Bousquet, 2018 and Zaffagni, 2018), but linguists were not involved in the revision process. Even though changes had to be implemented, the translation is considered to be of good quality, which is why this book is known to be the first book translated by an AI-powered system (Zaffagni, 2018). Consequently, it appears relevant to identify and analyse the instances in which the machine-translated text had to be edited.

For the purpose of this research project, the scientific director of Quantmetry accepted to provide the raw translation output of the third chapter, entitled "Probability and Information Theory" (pp. 51-76 in the English version and pp. 75-98 in the French version) which is approximately 9,000 words long, and was divided into 431 segments for this study. Therefore, the

[1] ENSAI: *École Nationale de la Statistique et de l'Analyse de l'Information* (National School of Statistics and Information Analysis)
INRIA: *Institut National de Recherche en Informatique et en Automatique* (National Institute of Research in Computer Science and Automation)
CNRS: *Centre National de la Recherche Scientifique* (National Centre of Scientific Research)

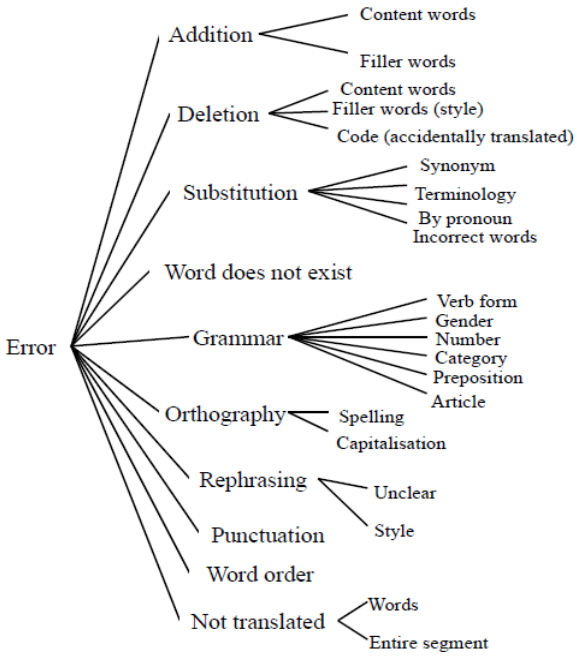

Figure 1: Error categorisation used for recording post-editing actions.

corpus consists of three texts – the original version of this chapter (English), the machine-translated text (French) and the published version (French).

3.2 Research Methods

Despite the popularity of automatic metrics such as BLEU, this research project is based on human evaluation, as it seems to be the most adequate method. Indeed, even though human evaluation is time-consuming and subjective, it allows for a more comprehensive classification of errors, and thus a more precise analysis of differences. Fundamentally, BLEU requires to have at least one reference human translation, which is not possible for this project. It could be argued that the post-edited version can be used as a reference. However, BLEU would still not be sufficient for the purpose of this research, in particular because it only focuses on the *n*-gram precision and it seems important to analyse larger units (as opposed to sequences of words).

Obviously, all of the taxonomies for human evaluation (mentioned in 2.2) were created with different purposes and are thus built differently. Error classification seems to be implemented on a case-by-case basis because the framework chosen to identify errors must be designed according to a number of characteristics. Two of the most important features are the type of text to be analysed and the languages involved (Vilar et al., 2006). However, previous studies do share some common features, as several categories appear to be recurrent (the major ones being the following: missing or additional words, incorrect words, word order, grammar, spelling and punctuation) and were therefore incorporated in the present research.

The material was analysed by a single bilingual annotator – the author of the present paper –, whose native language is French and who is entering the translation profession (one-year experience). The corpus was first gathered in a table with three main columns – the original text (English), the raw translation output and the post-edited, published version (French). Then, each segment of the corpus (i.e. sentences, titles, captions) was analysed manually and the changes spotted in the final version were recorded in a separate table. The classification adopted for recording errors was largely adapted from the error typologies proposed by Llitjós et al. (2005), Vilar et al. (2006), DePalma (2013), Lommel et al., (2015), TAUS Quality Dashboard (2016), Ahrenberg (2017), Daems et al. (2017), Hassan et al. (2018), and Popovic (2018). These are only a few examples of the studies presenting error typologies, and it is generally considered that the Multidimensional Quality Metrics (MQM) core (German Research Center for Artificial Intelligence, 2014) is a standard classification in the field (used in the study of Knowles et al., 2019, for instance). Consequently, the classification proposed here is also inspired by the MQM. The classification thus obtained is presented in Figure 1.

Errors were recorded by units of meaning. One error generally corresponds to one word edit (for example, a substitution edit corresponds in most cases to a single word edit). However, in the case that a post-editing action is affecting a unit of meaning composed of several words (e.g. “in terms of”), it was counted as one error.

Furthermore, a series of features above the sentence level was added to this classification. In particular, the instances in which sentences were split, merged, added or deleted were recorded. A particular emphasis was placed on the textual level, including in particular the consistency of terminology employed throughout the document

as well as coherence. Moreover, the translation procedures identified by Ahrenberg (2017) as being beyond the capacity of NMT (sentence splitting, shifts, explicitation, modulation and paraphrasing) were also studied.

4 Presentation of Results

Edit category	Edit sub-category	Total per sub-category	Total per category
Addition	Content words	63 (5.25%)	120 (10%)
	Filler words	57 (4.75%)	
Deletion	Content words	85 (7.08%)	148 (12.33%)
	Filler words (Style)	40 (3.33%)	
	Code	23 (1.92%)	
Substitution	Synonym	141 (11.75%)	389 (32.42%)
	Terminology	209 (17.42%)	
	By pronoun	12 (1%)	
	Incorrect words	27 (2.25%)	
Word does not exist		1 (0.08%)	
Grammar	Verb form	22 (1.83%)	238 (19.82%)
	Gender	76 (6.33%)	
	Number	58 (4.83%)	
	Category	24 (2%)	
	Preposition	33 (2.75%)	
	Article	25 (2.08%)	
Orthography	Spelling	0 (0%)	11 (0.92%)
	Capitalisation	11 (0.92%)	
Rephrasing	Unclear	36 (3%)	70 (5.83%)
	Style	34 (2.83%)	
Punctuation		57 (4.75%)	
Word order		19 (1.58%)	
Translation	Words in a segment	136 (11.33%)	147 (12.22%)
	Entire segment	11 (0.92%)	
Total number of edits		**1200**	

Table 1: Number of post-editing actions recorded per edit category presented as percentages.

Segments that did not require any edit	94 (21.81%)
Average number of edits per segment	2.78
Merged segments	4
Split segments	3
Added segments	1
Deleted segments	1
Untranslated segments	11 (2.55%)
Total number of segments in the corpus	**431**

Table 2: Segment analysis.

5 Evaluation

First, it should be pointed out that 21.81% of the output analysed did not require any edit, which proves that the NMT system was able to provide an output of publishable quality in certain cases. Moreover, the average number of edits per segment is 2.78, which corroborates the results obtained by Ahrenberg (2017).

While an error typology was used to easily record post-editing actions, it is deemed essential to point out that the types of edits identified belong to different severity levels. This concept was already used in previous studies, and it is the case in particular for the MQM, which relies on a scoring algorithm to assign a weight to the different errors encountered. Indeed, while untranslated words are obviously a critical issue, substituting a word by a synonym is a preferential edit (corresponding to the "preferential changes" master category in de Almeida, 2013) and thus belongs to a lower severity level.

5.1 Serious Errors

The most serious errors are attributed to cases of mistranslations. This happens when the raw output does not convey the meaning expressed in the Source Language (SL). A few words were translated incorrectly (2.25% of the edits performed), but they mostly correspond to bad translation choices and they generally do not interfere with the general meaning of an entire segment. Most of the time, errors belonging to the "incorrect words" category constituted a barrier to a good understandability and readability of the text because the formulation remained too close to that of the SL. At the sentence level, it was sometimes necessary to rephrase an entire clause or segment because the raw translation was not clearly formulated. Nevertheless, only 3% of the edits correspond to rephrasing an unclear segment in the output.

Other serious errors correspond to instances in which the output is not intelligible for the end reader. This obviously includes words that were left untranslated in the output. The post-editing action "translate" accounts for 12.22% of the changes made in the final version. In fact, 11/431 segments were not translated. It can be assumed that some words that were not translated are not commonly used, and since the material fed to the system did not contain instances of these words, it

could not translate them ("cassowary" for instance, appeared in English in the raw output). However, given that the segments that were not translated are not particularly challenging, it can be argued that this is due to a bug in the NMT system because this mostly happened in cases where the code surrounding these segments was particularly dense, in the case of captions for instance (6/11 occurrences).

Further serious errors include cases in which the output is not clear to the end reader. This happens in particular when the output presents a grammatical issue. For instance, the NMT system sometimes made errors of conjugation (especially regarding the sequence of tenses). On a few occasions, the grammatical category used in the output was not correct. However, it can also be the case that the sentence produced is grammatically correct, but the output is too close to the SL (literal translation) and therefore, the formulation does not seem natural to a French reader.

Moreover, even though a few serious grammatical errors were spotted in the NMT output, it is important to mention that most edits related to grammar were not implemented because the output was ungrammatical, but rather for the sake of consistency when other types of edits had to be performed. Indeed, changing even a single word in a sentence can have several repercussions and can thus considerably increase the number of edits necessary to produce a correct sentence, as already pointed out by Vilar et al. (2006). If a masculine noun was substituted by a feminine noun, it is likely that other elements in the sentence have to be modified (adjective and verb agreement, for instance). Nevertheless, the presence of acronyms resulted in recurring grammatical inconsistences, since acronyms seem to be identified as masculine by default in the NMT output.

One could think that segments in which it was necessary to add or delete words are severe cases of errors (especially for "content words"), but most often, this did not affect the general meaning of a segment. These changes are sometimes preferential, and in other cases, words were added in order to make the target text more precise or when it was deemed necessary to include additional information. When it comes to deleting words, this could be done when a concept was implied or simply for stylistic reasons (to avoid repetitions in particular). In fact, the most serious case in which words needed to be deleted was when the code was accidentally translated (for example: "**\newterm{**multinomial distribution**}**" translated as "*nouvelle distribution multinomiale"* [new multinomial distribution], instead of "*distribution multinomiale"* [multinomial distribution]).

Furthermore, only one word that does not exist was spotted in the output ("*prioror*", which resembles both the English "prior" and the French "*à priori*").

As far as orthography is concerned, spelling mistakes were included in the classification, but none was found in the raw output. Only a few capitalisation errors were spotted, accounting for 0.92% of the edits performed.

5.2 Contextual Errors

The following severity level corresponds to errors related to the context. In fact, this is the case for most of the errors in the output analysed. Most of the output was grammatically correct and understandable, but the lexical items employed needed to be adjusted to comply with terminological standards. The specialised terminology apparently constituted a genuine challenge: 32.42% of the edits were substitutions, and 54% of the substitutions were performed to comply with terminological requirements. In most cases, specialised terms were not identified and were translated as general words, which is a rather unexpected finding, given that a glossary of specialised terms was used for the translation of *Deep Learning*. Furthermore, on a few occasions, some inconsistencies in terminology were spotted in the NMT output.

It is also essential to point out that, even though substitutions performed because of the specialised terminology correspond to the most common type of edit, several errors were in fact replicated but recorded as often as they appeared in the raw MT output. As a result, an important number of changes were performed to correct the same error appearing multiple times. This is particularly true for this error category and thus contributed to make it the most prominent in the results.

5.3 Stylistic Preferences

The last level of severity corresponds to preferential changes. These edits were not performed to correct grammatical or terminological errors, but are rather based on

stylistic preferences. In particular, the authors are clearly present in the SL, which is particularly reflected by the use of personal pronouns ("we provide this chapter to ensure…", p.51 of *Deep Learning*). This is not common in French: Pontille (2006) underlined that markers of the authors' presence should be carefully erased in scientific texts in order for the readers to focus on the facts presented. This element was thus modified at the post-editing stage in order to make the French version more impersonal. Other preferential changes include instances in which a noun was substituted by a synonym (11.75%) as well as reformulation of a sentence based on stylistic preferences (2.83%).

Moreover, even though most standards of scientific writing encourage repetition in French (for instance, Boudouresque, 2006) for the sake of precision, Baker (2018) pointed out that the acceptability of this procedure varies greatly across languages. In fact, even in scientific discourse, French generally tends to avoid repetition in order to enhance readability. For this reason, pronouns were used in the final version of the text ("substitution by pronoun"). Alternatively, some words could be deleted when they were not deemed necessary or when they were mentioned shortly before.

5.4 Procedures Beyond Reach of NMT

The outcomes of this study confirmed the observations made by Ahrenberg (2017) regarding the translation procedures beyond reach of NMT systems. No sentence was split in the raw output. On one occasion, two sentences were merged in the raw output, which demonstrates the ability of NMT technology to handle sentences, but this would need to be analysed in more detail. Modulation and explicitation also appear to be beyond the capacity of NMT. Similarly, category shifts and paraphrasing seem to be procedures that the NMT system did not implement, which sometimes caused the output to be too literal. In addition to these procedures, it appears important to mention that the NMT system was not capable of making adjustments regarding the readability (e.g. substitution by pronoun to avoid repetition) and the register (some sentences were translated literally and would certainly have been acceptable in oral discourse, but needed to be changed to a more formal tone).

6 Limitations of the Research

The first limitation of this research corresponds to the size of the corpus analysed (only one chapter of *Deep Learning*). Even though the chapter analysed can be considered as representative of the entire book, verifying whether the results obtained in this study apply to the whole text would certainly constitute a valuable analysis. Beyond a larger sample of the same book, it would also be relevant to extend this study to different text genres in order to verify whether it would show similar results.

The same goes for the linguistic combination. This research project only focused on the English-French language pair, whereas several NMT systems offer a number of different combinations. It would therefore be relevant to evaluate NMT output for more distant languages. This could help in identifying strengths and weaknesses of such technology that are independent of the language pair studied.

Another limitation lies in the MT system itself. Indeed, Quantmetry has developed a NMT tool in partnership with DeepL GmbH for the purpose of translating *Deep Learning* into French. It was announced at the DataJob conference (2018) that the company aimed at making this tool available for free to the public in the months following the publication of the book in France, but it was not the case by the time this research project was conducted. Nevertheless, after comparing some fragments of the raw translation obtained and a translation of the same text performed by DeepL's online NMT tool (excerpt of 40 segments), it seemed that both outputs were particularly similar (about 80% of the segments tested were identical), which is understandable given the contribution of DeepL GmbH to this project. Therefore, the data analysed in this research project can be considered as representative of NMT output. However, conducting human evaluation on more NMT systems would allow to verify whether the results obtained in the present study are applicable to more NMT systems.

The methodology adopted for this research also constitutes a limitation. The classification implemented for the analysis of the corpus was inspired by previous studies, and only the features that seemed relevant to this project were selected. One can argue that a more exhaustive typology should be built, thus allowing to analyse more

aspects in future projects. Beyond this point, establishing a new error typology makes the experiment hardly reproducible and comparable to other research in the same area.

Moreover, this research project is based on human evaluation only, and relying on a single annotator compromises the analysis. Hence, a suggestion for further research would be working with more evaluators, as inter-rater agreement testing is particularly valuable when assessing post-editing. It would also appear relevant to analyse the same corpus with automatic metrics and to compare the results thus obtained with the findings of the present study, as already suggested in the research of Vilar et al. (2006).

Finally, it must be acknowledged that the present research would have benefitted from more information regarding the NMT system used for the translation of *Deep Learning* (i.e. training methods used, post-editing guidelines followed, etc.). Unfortunately, this information was not available by the time the present research was conducted.

7 Conclusion

This research project provided a detailed analysis of the changes performed at the post-editing stage in the case of *Deep Learning*. The interpretation of the results obtained allowed to meet the objective of this project by identifying recurring patterns of errors, thus providing an evaluation of the raw NMT output.

What emerges from this study is that the NMT tool produced critical errors in some instances, but several changes made in the final text were preferential and the majority of edits were performed to comply with terminological standards. Of course, this evaluation is largely subjective, and the raw NMT output would not have been acceptable without any post-editing. But it seems reasonable to say that the NMT tool developed for the translation of *Deep Learning* was efficient and that the raw translation is satisfactory for the intended use of this kind of material nowadays, knowing that machine-translated texts are still reviewed.

The evaluation of NMT conducted for this research provides translation professionals and scholars with an insight of the performance of NMT in the case of *Deep Learning* as well as a list of predominant errors in NMT, which correspond to aspects that should be carefully controlled at the post-editing stage in the English-French combination.

As things stand currently, NMT tools are still not efficient enough for producing translations of human quality, as the raw output analysed in this project is not comparable to a human translation. Nevertheless, artificial neural networks are a very promising technology and with the increasing amount of data produced, NMT seems to be an ideal solution to meet the translation demand. But even in this scenario, human translators will play a key role, as the development of more efficient MT tools will mostly depend on collaboration between computer engineers and professional translators. Therefore, it seems essential to implement an 'orchestrated symbiosis' (in the words of Bawa-Mason et al., 2018); it is crucial that translators do not consider technology as a competitor but as a means to enhance their performance. Working hand in hand with computer engineers is essential to improve MT systems. Such collaboration would allow engineers to understand better the equivalence issues between languages as well as typical translation problems and thus to design new systems able to provide even better results.

The analysis conducted for this project provides a list of features that NMT specialists should endeavour to improve when developing new tools (language in context, the importance of specialised terminology, etc.). Furthermore, receiving feedback from linguists working with NMT systems is also essential for the implementation of more sophisticated automatic metrics suitable for the evaluation of more recent MT tools.

Acknowledgments

I would like to express my sincere gratitude to Nicolas Bousquet. I would not have been able to conduct this research project without his help and the raw translation output he kindly accepted to provide me.

References

Lars Ahrenberg. 2017. Comparing Machine Translation and Human Translation: a Case Study. In *RANLP 2017 The First Workshop on Human-*

Informed Translation and Interpreting Technology (HiT-IT): Proceedings of the Workshop, ACL, Shoumen, Bulgaria, pages.21-28. https://doi.org/10.26615/978-954-452-042-7_003.

Mona Baker. 2018. *In Other Words: A Coursebook on Translation*. Routledge, London, UK, 3rd edn.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics*, ACL, Ann Arbor, USA, pages 65-72. https://www.aclweb.org/anthology/W05-0909.

Sarah Bawa-Mason, Lindsay Bywood, Charles Gittins, Paul Kaye, Raisa McNab, Maeve Olohan and Michael Wells. 2018. Translators in the Digital Era: What Kind of Jobs Will We Have Ten Years from Now? Presented at *The Language Show*, Olympia, London, UK, 11 November 2018.

Charles-François Boudouresque. 2006. Manuel de Rédaction Scientifique et Technique. Centre d'Océanologie de Marseille. 4th edn.

Nicolas Bousquet. 2018. Deep Learning: Histoire d'une Traduction. Presented at *Data Job*, Carrousel du Louvre, Paris, France, 22 November 2018.

Chris Callison-Burch, Miles Osborne and Philipp Koehn. 2006. Re-Evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Trento, Italy, pages 249-256. https://www.aclweb.org/anthology/E06-1032.

Joke Daems, Sonia Vandepitte, Robert J. Hartsuiker and Lieve Macken. 2017. Identifying the Machine Translation Error Types with the Greatest Impact on Post-Editing Effort. Frontiers in Psychology, vol. 8, 1282. https://doi.org/10.3389/fpsyg.2017.01282.

Giselle de Almeida. 2013. *Translating the Post-Editor: an Investigation of Post-Editing Changes and Correlations with Professional Experience across Two Romance Languages.* PhD thesis, Dublin City University.

Don DePalma. 2013. *Post-Editing in Practice*. TCworld.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (German Research Center for Artificial Intelligence). 2014. *Multidimensional Quality Metrics (MQM) Definition*. http://www.qt21.eu/mqm-definition/definition-2015-06-16.html

Gino Diño. 2018. 'Human Parity Achieved' in Machine Translation — Unpacking Microsoft's Claim. *Slator*.

Thi Ngoc Diep Do. 2011. *Extraction de Corpus Parallèle pour la Traduction Automatique Depuis et Vers une Langue Peu Dotée*. MSc Thesis, Université Grenoble Alpes, France; Hanoi University, Vietnam.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics. In *Proceedings of the Second Conference on Human Language Technology*, ACL, San Diego, USA, pages 138-145. https://doi.org/10.3115/1289189.1289273.

Ian Goodfellow, Yoshua Bengio and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge.

Ian Goodfellow, Yoshua Bengio and Aaron Courville. 2018. *L'Apprentissage Profond*. Florent Massot, Paris, France.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. Microsoft AI & Research. https://www.microsoft.com/en-us/research/uploads/prod/2018/03/final-achieving-human.pdf.

Basil Hatim and Jeremy Munday. 2004. *Translation: An Advanced Resource Book*. Routledge, London, UK.

Pierre Isabelle, Colin Cherry and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, ACL, Copenhagen, Denmark, pages 2486-2496. https://aclweb.org/anthology/D17-1263.

Rebecca Knowles, Marina Sanchez-Torron, and Philipp Koehn. 2019. A User Study of Neural Interactive Translation Prediction. *Machine Translation Journal Special Issue on Human Factors in Neural Machine Translation*. vol. 33, pages 135-154.

Alon Lavie. 2011. Evaluating the Output of Machine Translation Systems. Presented at the *13th MT Summit Tutorial*, 19 September 2011, Xiamen, China.

Samuel Läubli, Rico Sennrich and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In

Proceedings of EMNLP, ACL, Brussels, Belgium pages 4791-4796. https://aclweb.org/anthology/D18-1512

Ariadna Font Llitjós, Jaime G Carbonell and Alon Lavie. 2005. A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*, EAMT, Budapest, Hungary, pages 87-96. https://pdfs.semanticscholar.org/ba58/aa555d6be8fd5b5c148ff3daf992c3a1803d.pdf.

Arle Lommel, Attila Görög, Alan Melby, Hans Uszkoreit, Aljoscha Burchardt and Maja Popović. 2015. QT21 Harmonised Metric. QT21 Consortium.

Johanna Monti, Violeta Seretan, Gloria Corpas Pastor and Ruslan Mitkov (eds). 2018. *Multiword Units in Machine Translation and Translation Technology*. John Benjamins, Amsterdam and Philadelphia.

Hazel Mae Pan. 2016. How BLEU Measures Translation and Why It Matters. Slator.

Ramón P. Ñeco and Mikel L. Forcada. 1997. Asynchronous Translations with Recurrent Neural Nets. In *Proceedings of the International Conference on Neural Networks*, IEEE, Houston, TX, USA, vol. 4, pages 2535-2540.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL, Philadelphia, USA, pages 311-318. https://www.aclweb.org/anthology/P02-1040.

Emmanuel Planas. 2017. État de l'Art et Futur de la TAO en 2017. Presented at *Penser la Traduction 2017-2018*, Faculté des Lettres, Langues et Sciences Humaines de l'Université de Haute Alsace, France, 27 November 2018.

David Pontille. 2006. Qu'est-ce qu'un Auteur Scientifique? *Sciences de la Société, Presses universitaires du Midi*, pages 77-93. https://halshs.archives-ouvertes.fr/halshs-00261793/document

Maja Popovic. 2018. Error Classification and Analysis for Machine Translation Quality Assessment. In *Translation Quality Assessment*. pages 129-158. https://www.researchgate.net/publication/325896250_Error_Classification_and_Analysis_for_Machine_Translation_Quality_Assessment.

Rico Sennrich. 2016. *Neural Machine Translation: Breaking the Performance Plateau*. Presented at the META-FORUM 2016, Lisbon, Portugal. http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

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 the 7th Conference of the Association for Machine Translation in the Americas: Visions for The Future of Machine Translation*, AMTA, Cambridge, Massachusetts, USA, pages 223-231. https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf.

TAUS Quality Dashboard. 2016. From Quality Evaluation to Business Intelligence. TAUS BV, De Rijp, The Netherlands.

Christoph Tillmann, Stephan Vogel, Hermann Ney, A. Zubiaga and Hassan Sawaf. 1997. Accelerated DP-Based Search for Statistical Translation, In *Proceedings of the 5th European Conference on Speech Communication and Technology*, EUROSPEECH 1997, Rhodes, Greece, pages 22-25. https://pdfs.semanticscholar.org/2472/ad58de68c65d54e05470cceee70b4f4f8bb3.pdf.

Antonio Toral, Sheila Castilho, Ke Hu and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation*, ACL, Brussels, Belgium, pages 113-123. https://www.aclweb.org/anthology/W18-6312

Joseph P. Turian, Luke Shen and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of Machine Translation Summit IX*, AMTA, New Orleans, LA, USA, pages 386-393. https://nlp.cs.nyu.edu/publication/papers/turian-summit03eval.pdf.

Ilya Ulitkin. 2013. Human Translation vs. Machine Translation: Rise of the Machines. Translation Journal. vol. 17. https://translationjournal.net/journal/63mtquality.htm.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Statistical Machine Translation Output. In *Proceedings of the International Conference on Language Resources and Evaluation 2006*, Genoa, Italy, pages 697-702. http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf.

Marc Zaffagni. 2018. Une IA Traduit un Livre de 800 Pages en 12 heures. Futura Tech.