Automatic Classification, Visualization and Analysis of Errors in Machine Translation

Chathuri Jayaweera
University of Moratuwa
chathurij@uom.lk

Gihan Dias
University of Moratuwa
gihan@uom.lk

***Abstract*—Although the quality of machine translation (MT) has improved in recent years, machine translated documents still contain errors. MT quality is often evaluated using a single numeric score. However, this may not adequately characterise the system. We provide an error visualizer, which shows differences between corresponding lines of two translations. In addition to insertions, deletions and substitutions, our system also shows *transpositions*. We also provide an error analyzer which gives statistics of each type of error in the document. In addition, it shows errors in *context*: the words commonly adjacent to each error, and also the adjacent parts of speech (POS). This feature - unique to our system - allows the identification of the context in which errors occur, so they can be rectified easily. The system was evaluated by three MT system developers, who identified useful features and provided feedback which was used to improve the system.**

***Index Terms*—comparison, error analysis, error classification, evaluation, machine translation, MT**

I. Introduction

Although the quality of Machine Translation (MT) has improved in recent years, many errors are still observed in a machine translated document. These errors may be categorized as deletions, insertions, re-orderings (transpositions), substitutions etc. when compared with a reference translation. These errors reduce the accuracy of an MT system.

Analysing errors is important to identify error patterns and the cause of errors and in developing with solutions to reduce them. Numerical scores for evaluation of machine translation such as BLEU, METEOR etc. only give an idea of the performance of an MT system and may not align with human evaluations. Further, these scores do not provide useful insights on errors that can be gained through manual comparison and analysis. However manual analysis is time-consuming, tedious and expensive. Therefore, a system that automatically classifies and qualitatively analyzes errors is useful in developing MT systems. This paper discusses our attempt in implementing an analysis tool that meets the above requirements.

II. Related Work

A number of software and tools are available for automatic error classification, analysis and evaluation. ibleu [1], MTEval [2] and Vis-Eval [3] use automatic evaluation metric scripts.

MT-ComparEval [4] is an interactive user interface that provides both quantitative and qualitative insights on multiple machine translation models such as pair-wise comparisons of MT systems, statistical insight on sentence-level scores etc. Compare-mt [5] provides holistic comparisons and analyses on two MT systems' performance based on several components of the translations such as sentence length, translation accuracy etc. VizSeq [6] is a visual analysis toolkit for language generation tasks. Hjerson [7] is an automatic error classification tool for machine translation outputs which provides statistics on each type of error identified.

III. Methodology

We utilized a Sinhala - English parallel corpus of 1603 sentence pairs from government documents and websites for this analysis. Three MT systems, namely *Google Translator*, A fine-tuned *MBart* model and a *Transformer* model were used to translate the Sinhala source sentences to English. These outputs were compared and analysed to identify their respective error patterns and performance in handling different error types.

Our system comprises two parts:

A. sentence-wise visual comparator between two texts

We performed a visualisation of errors in each sentence in each MT text with respect to the reference translation and marked each erroneous word with its respective error type. This may be visualized in a terminal or a Jupyter notebook.

B. statistical analysis of errors

We analyse errors for each type of error: deletion, insertion, substitution and re-ordering/transposition. In single-word frequency analysis, the frequency of each erroneous word is calculated. E.g., if ”the year” is deleted, this is treated as *two* deletions, of ”the” and ”year”. In multi-word analysis, contiguous error words are considered a single error. E.g., deletion of ”the year” is considered a single error.

In contextual analysis, the frequency of each erroneous word is calculated in the context of (i) the preceding and (ii) the succeeding word (bigram frequency). We also perform a tri-gram frequency analysis of each erroneous word in the contexts of the POS tags considering preceding and succeeding words together with the erroneous word. This may provide better insight than using word bigrams. We also calculate the trigram frequencies of each error word in the context of its preceding and succeeding POS tags.

IV. Results and Analysis

A. Error Classification and Erroneous words identification

Through our analysis, we found that the most frequent words deleted (i.e, in the reference text but not in the MT) are determiners, auxiliary verbs, prepositions, etc. In the single-word mode, only the word "year" did not belong to the aforementioned categories. In the multi-word mode, we observe several words and phrases such as "Name / RDA Division", "Mr.", and "the year" that have been deleted frequently. From the frequencies of the deleted words, it was observed that *MBart* translation performs better than the other two systems for deletions, but it was observed that *Transformer*'s performance in multi-word deletion is better than both *GT* and *MBart* models. In terms of insertion errors, in the single-word mode, the *MBart* model performs poorly for some of the erroneous words. However, *GT* inserts "The" more often than the other systems.

GT performs better in handling re-ordering of words than the other two systems. However, when it comes to "and" and "Rs", *GT* shows a higher occurrence of re-ordering. We analysed the nature of the substitutions that have occurred in the translations. It was observed that in most of the instances, only a difference of a capital letter at the beginning of the word had occurred, and in some cases, abbreviations were substituted by their extended terms (eg: "AO" with "Accounting Officer", "SL" with "Sri Lanka" etc.). For these instances, most of the substitutions that have occurred through these MT systems do not significantly affect the meaning of the sentences, but automatic evaluation metrics that check for n-gram sequence matches penalize them. Another significant observation is that *GT* has replaced "Km" with "km". This shows that SI units were not properly used in the reference translation but *GT* has translated them properly. We observed that *MBart* has learned an incorrect translation for that unit as there are "km" to "Km" substitutions by *MBart*. Therefore, it is clear that through this analysis, we can identify not only patterns in MT system errors but also errors in the reference translation.

B. Common contextual errors

We performed a contextual analysis for each erroneous word of each error category. We observed that for all the three MT systems "IN DT NN" which corresponds for "preposition+determiner (eg: "the")+noun" and "IN DT NNP" which corresponds for "preposition+determiner ("the")+singular proper noun" are the most frequent context for the word "the" to be inserted. The next most frequent context is "ID DT JJ" which corresponds to "preposition+determiner ("the")+adjective". When this third sequence was further analysed, we observed that the word "the" has been inserted in contexts such as "in the Southern province", "of the Western province" etc. corresponding to this POS tag sequence. In this way, we can retrieve insights on most common contexts for a particular erroneous word to be inserted, deleted, replaced or transposed through contextual error analysis.

V. Conclusion and Future Work

The error visualiser allows an MT system developer to analyse the differences either between lines of a reference and a machine translation, or between two MT systems (or versions). Colour-coding allows the prevalence of each type of error to be ascertained at a glance, and the errors in each line to be identified. The analysis of each error in its context, and in using POS-based context, are unique features of our system. We did not perform a quantitative analysis of the system, but qualitative feedback shows that the system is useful in debugging MT systems.

At the highest level, our error analyser calculates a set of metrics (frequencies of insertions, deletions, substitutions, transpositions, etc.) for an MT system, rather than just a single score. At the non-contextual level, the system provides statistics on each word in error. The feedback received showed that words which are commonly in error (e.g. determiners) are identified. At the contextual level, we identify the common adjacent words for each word in error. We may then augment our training data with real or synthetic data to avoid such errors.

We identified that some so-called errors were actually due to errors in the reference translation. The machine translation was actually better than the reference. We then corrected the reference data, and plan to re-train the MT system with the corrected training data.

The contextual error analysis currently only looks at adjacent words and POS tags. We may include more types of adjacency, and also consider nearby (non-adjacent) words.

We have developed an error visualiser and analyser which is useful for machine translation developers. It has been successfully used to improve the performance of Sinhala → English MT systems, but may easily be adapted to other language pairs.

References

[1] N. Madnani, "ibleu: Interactively debugging and scoring statistical machine translation systems," in *2011 IEEE Fifth International Conference on Semantic Computing*. IEEE, 2011, pp. 213–214.

[2] A. Bharati, R. Moona, S. Singh, R. Sangal, and D. M. Sharma, "Mteval: an evaluation methodology for machine translation systems," in *Proc. SIMPLE Symp on Indian Morphology, Phonology and Lang Engineering*. Citeseer, 2004.

[3] D. Steele and L. Specia, "Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018, pp. 71–75.

[4] O. Klejch, E. Avramidis, A. Burchardt, and M. Popel, "Mt-compareval: Graphical evaluation interface for machine translation development," *The Prague Bulletin of Mathematical Linguistics*, vol. 104, no. 1, pp. 63–74, 2015.

[5] G. Neubig, Z.-Y. Dou, J. Hu, P. Michel, D. Pruthi, X. Wang, and J. Wieting, "compare-mt: A tool for holistic comparison of language generation systems," *arXiv preprint arXiv:1903.07926*, 2019.

[6] C. Wang, A. Jain, D. Chen, and J. Gu, "Vizseq: A visual analysis toolkit for text generation tasks," *arXiv preprint arXiv:1909.05424*, 2019.

[7] M. Popović, "Hjerson: An open source tool for automatic error classification of machine translation output," *The Prague Bulletin of Mathematical Linguistics*, vol. 96, pp. 59–67, 2011.