# ANLE Report:
## Exploration of Machine Translation Approaches

Candidate Number:184521

## Introduction

Machine Translation, MT, is an automated translation process of converting one Natural Language to another, for example English to Spanish, by using a computer algorithm and is one of the many areas of research found in "Computational Linguistics". Two important distinctions to be made are that the text that is translated can be either written or spoken and that the aim of MT is to restore the original meaning of the text in the translated text [1][2][3]. Real world examples of this include Google Translate, used by the public for day to day translations, or the professional translation industry who use machine translation to increase productivity of workplaces [4].

The MT process usually consists of two different components, the metaphrase and paraphrase, with metaphrase relating to the word to word translation, or formal equivalence, whilst the paraphrase relates to the dynamic equivalence of the text. The former however may not give the original meaning of the text as it is simply a literal translation of each word unlike the latter which translates the general context of the text but not the individual words [3]. This process has stayed largely the same, but the approaches have evolved from Rule based to Neural since its conception in the 1940's. Figure 1 shows how these processes can take different routes allowing for different methodologies with the most basic being direct word to word translation with this losing all syntactical, semantic and morphological information. The most complex method would be to go Interlingua with a complete decomposition/analysis and regeneration of each element which is something that the different MT methods try to complete.[1][2]

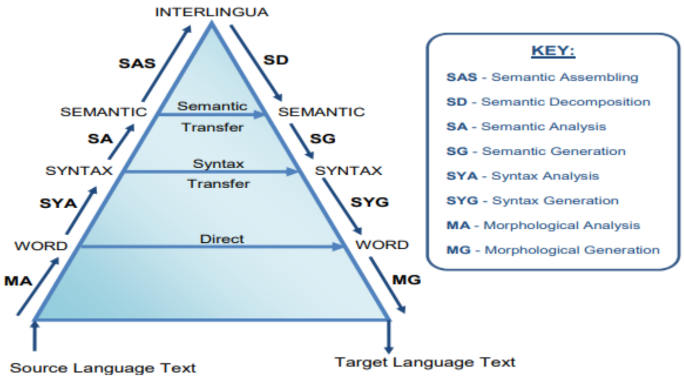

Figure 1: Vauqouis triangle [1]

This is as MT is a difficult process to complete meaning changes in its methodology has looked to reduce these difficulties which is one of the many reasons the approaches have evolved from Rule-Based to Neural. However, it is debated whether the new approaches do indeed work better than the original Rule-Based approach and it's this comparison that is explored in this report.

One of the many difficulties that faces MT is the structural differences between the original text and the translated text whether these are morphological differences or syntactic differences [5]. Morphological differences involve the differences in the components, morphemes, of the words such as the Prefixes and Suffixes of a word. Syntactic differences on the other hand is more about the grammar and order of words in the given origin text. Syntactic differences are not easy to replicate between languages as, for example, English puts its adjectives before a noun whereas a language such as French will put them after the noun.

Another difficulty the MT is required to solve are lexical differences between one language and another [5]. This is how languages diverge in the methods they use to form sentences with languages such as Spanish and French having different genders for their words but languages such as English not. These divergences need to be accounted for when translating between languages. Researchers in the field of Computational Linguistics have developed solutions for these difficulties with the study of linguistic typology, where they study the differences and similarities between different languages. One of these solutions are rough translations for providing the underlying context of the original text to the translated text allowing the final text to contain the same information as the original. Another solution is post editing in which humans aid the computers with the final translation by recognising grammatical or syntactical errors and replace them for the computer, though this is not ideal for an automated solution [6]. A final proposed solution is that of domain restrictions, which is utilised by Neural approaches, where a domain displays the context or category the text is about. By restricting the domains only words related to certain domains can be used in the translated solution.[7][5] Therefore over the duration of the report the original Rule based and newer neural approaches shall be compared on technical and performance based levels.

## Rule-based Machine Translation (RBMT)

The Rule Based approach, RB, to MT revolves around the morphological, syntactical, and semantic information of the given original language and the target translation language. This means that the RB approach looks at collections of grammar rules and structures of the different languages and uses this to convert one to the other [2][3]. In doing so multiple bilingual dictionaries are utilised to give accurate translations allowing the context to be easily maintained. One downside with the rule-based approach is that it has trouble applying grammar exceptions, an example of which is the common ‘I before e except after c’ [3]. This approach also heavily uses language theory, which requires many hours of human labour to correctly build the rules to be used by this approach. However due to these rules being easily manipulated by a human this allows for maintenance and the addition of more languages to be done with ease although the former must be done regularly to keep accuracy up. This models effectiveness can also easily be measured by its fluency, precision, post-edit, and fidelity which all show a generally good performance.

## Neural Machine Translation (NMT)

Neural Machine Translation, NMT, involves using different neural network models to learn statistical matrices used during Statistical MT with the goal of translating one language to another without any extra input. This is done by neural networks learning the morphological, syntactical, and semantic similarities between one language’s texts and another using the Statistic Matrix [10]. However, the performance of the NMT can vary between implementation as different matrices and Neural Networks can be used. For example, a basic architecture for an NMT would be to use two Recurrent Neural Networks, RNN, where one is used to consume input text, and another is used to generate the translated text [5]. However different Neural networks could be used instead depending on the complexities and differences between languages as some would perform better than others. For example, RNNs could be replaced with CNNs, Concurrent Neural Networks if sentences had a higher complexity as these would give a better result as shown by Tan et al [11].

# Performance and Evaluation

When reviewing the performance of the RB and NMT approaches the best method to use would be automatic evaluation using the BLEU evaluation metric, which is an algorithm used to evaluate the quality of the translated text given by a piece of MT software. Where the main principal is the closer to a human translation the software's output is the better its performance and is conducted by checking the correlation of a software's output with that of a human [9]. Due to this the BLEU score is not 100% effective for comparing the approaches as it mainly measures direct word-to-word and word cluster similarity. However, no papers comparing NMT and RB approaches were found using this methodology.

Due to this Macketanz et al's 'Phrase-Based, Rule-Based and Neural Approaches with Linguistic Evaluation' paper was used [12]. This paper uses multiple evaluation methods to gain insights into the strengths and weaknesses of each of the approaches by seeing how they compare comparatively, with this paper utilising bidirectional RNNs in its NMT construction. The first evaluation that was conducted was via the Manual Linguistics method which involved a German linguist evaluating the translated texts outputted by the different MTs for errors. This was then used to obtain the accuracy of each approach. In these circumstances the RB approach obtained a 76% accuracy whereas the NMT was at 75% giving only a 1% difference between the two. More in depth information was given on the accuracy of the different phenomenon used in the two languages. For instance, the RBMT seemed to perform better with Imperatives, compounds and phrasal verbs with the latter having a 67% accuracy, 29% higher than the NMT. The NMT on the other hand performed better with terminology, quotation marks and ">" separators with the latter here being 66% higher accuracy than RBMTs. It should be noted other than ">" separators, compounds, and phrasal verbs all other accuracies are within 5% of each other, with most of them being in the RBMTs favour.[12]

The next experiment performed by Macketanz et al was the evaluation of test suite data where 100 test sentences placed in 14 different categories were fed to the MTs with the evaluation of the translated sentences being performed in the same way as the previous experiment. Here RBMT achieved an average 69% accuracy whereas neural only achieved 48% accuracy. This is with RBMT excelling in categories of composition, non-verbal agreement, and verb tenses with accuracies ranging from 100% to 7% better than that of the NMT. However, the NMT did perform better in function word and punctuation categories by 25% and 50% respectively, though these were the only categories out of the 14 that NMT performed better in. This trend is replicated in Burchardt et al's paper [13] where multiple NMT methods using different Neural Nets are trialled against RBMT with RBMT getting 83% whilst the NMT methods accuracies range between 50-76%. These methods used similar size corpus to Macketanz et al replicating how well RBMT's perform on small amounts of data.

Due to these results, the RB approach seems to be the most effective, especially when only small corpuses are available, making it the best choice for use when a language has limited resources in which to create its MT software. This is as it is alluded to that the corpus used for these experiments was not that large and therefore RBMT excels at simple translations which don't require much material to calculate [9]. Though this is true, due to the high human resource element of RBMTs the NMTs are more widely used and as more advanced computers become available the required computational power to use an NMT effectively will become lessened. Also in the real world larger corpuses are used to train the NMTs with smaller corpuses being used in these experiments being a time saving method. This is as large corpuses would increase the time it takes to train an NMT and create rules for the RBMT. This is on top of only a small selection of NMTs methods were trialled here with other methods having the possibility of performing better.

The main benefit of the NMT approach over others is that it is quick and easy to train only requiring the original text and models used for training making this a complete end-to-end system [10]. However, as soon a limitation of this is the size of the corpus used in training the NMT as it must be sufficiently large for the NMT to perform correctly. This differs to that of Rule-Based MT where a human must manipulate and set up the grammatical rules required for the system to work which requires maintenance and post-editing to keep correct. Although this is easy to perform it does increase the time and human cost of running the particular MT. Therefore, in approaches NMTs are more streamlined than RB due to it being an end-to-end solution allowing it to be fully automated without any human input needed. One of the reasons for this is the use of domain restriction to reduce the difficulty of the task being done. As stated previously, domains are categories that text can fit into and by restricting these the context of the text can be maintained by not allowing the neural network to change it from one domain to another [7].

# Conclusion

In conclusion, with the small corpus used in the Macketanz et al's paper and others the RBMT is shown to perform better in most categories with higher accuracies given to them. However, with the requirements needed to set up an RBMT it is possible to see why NMTs are becoming the norm with companies such as Google switching to use them. This is as with large enough corpus's and computation power they can become effective. However, this is something that has not been researched in depth. Therefore, the hypothesis that if given enough data different methods of NMTs will give better accuracies then RBMTs should be tested in a further study.

# Bibliography

Word count:2164

Bibliography word count: 132 words

[1] Machine Translation Approaches: Issues and Challenges https://www.ijcsi.org/papers/IJCSI-11-5-2-159-165.pdf

[2] Review on Machine Translation Approaches https://www.researchgate.net/profile/Benson-Kituku/publication/299435025_A_Review_on_Machine_Translation_Approaches/links/56f94f8908ae95e8b6d3fc6a/A-Review-on-Machine-Translation-Approaches.pdf

[3] Approaches to machine translation http://nopr.niscair.res.in/bitstream/123456789/11057/4/ALIS%2057%284%29%20388-393.pdf

[4]Who use machine translation https://omniscien.com/faq/who-uses-machine-translation/

[5]Machine translation lecture https://canvas.sussex.ac.uk/courses/12968/pages/week-slash-topic-10-machine-translation-1

[6] Post-editing Machine Translation https://www.trados.com/learning/training/post-editing-machine-translation.html#:~:text=Post%2Dediting%20means%20that%20a,learning%20curve%20associated%20with%20PE.

[7] Domains and domain restriction; Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. CoRR, abs/1808.10432.

[8] CRITICAL LOOK AT THE PERFORMANCE OF RULE-BASED AND STATISTICAL MACHINE TRANSLATION; https://www.scielo.br/scielo.php?pid=S2175-79682020000100054&script=sci_arttext

[9] a Method for Automatic Evaluation of Machine Translation: https://www.aclweb.org/anthology/P02-1040.pdf

[10] NMT introduction https://machinelearningmastery.com/introduction-neural-machine-translation/

[11] Neural machine translation: A review of methods, resources, and tools

https://www.sciencedirect.com/science/article/pii/S2666651020300024

[12] Phrase-Based, Rule-Based and Neural Approaches with Linguistic Evaluation https://www.dfki.de/~ansr01/docs/MacketanzEtAl2017_CIT.pdf

[13] Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines: http://archive.sciendo.com/PRALIN/pralin.2017.108.issue-1/pralin-2017-0017/pralin-2017-0017.pdf