**Capstone Project: Machine Translation**

**Milestone2: Project Synopsis**

**CONTENT**

1. **Problem Definition**
2. **Literature Survey**
3. **Data**
4. **Tentative Roadmap**
5. **References**
6. **Problem Definition**

Communication is the key for any business, education, networking or any type of activity requires for human existence. And the medium to communicate with any person is human language. Whenever we buy an international product the instruction manual contains different languages. In ancient times, when reachability was very limited to locality due to lack of means of transportation and media, many civilizations came up with their own languages. This leads to linguistic diversity that we have today in the world. The most extensive catalog of the world’s languages, generally taken to be as authoritative as any, is that of [Ethnologue](http://www.ethnologue.com/) (published by SIL International), whose detailed classified list as of 2009 included **6,909 distinct languages**. In India alone, linguistic diversity is astounding. The magnitude of language diversity in India can be gauged from the fact that according to the Census of India (2011), **there were 121 major languages in India with 1599 other languages.** Sanskrit and Tamil are considered some of the ancient languages in the world that came from India.

However, linguistic diversity creates many obstacles, considering the number of languages present across the world and capability of humans to completely grasp only few, given the limited efforts they can put in this activity in lifetime. Learning a language is not an easy task. Every language is unique in terms of lexical, structural and/or semantic properties. Scripting, phonetics, dialect and grammar – everything changes from one to another. This confines human connectivity with in the linguistic boundary. Often due to language barrier a company/individual is unable to sell any product or make a deal. Tourists face issues while traveling other countries. Other challenge is that languages evolve over the time. Many ancient texts are still untranslated, which is just loss of meaningful information related to human lives. Therefore, If the language barrier can be resolved, then communication will happen at a larger scale and the outreach will be more, thus enhancing connectivity, economy, knowledge sharing and so on.

Humans might have physical limitation but no intellectual limitation. Man-made Machines have always been making lives comfortable to achieve something which is otherwise challenging. Machine Translation (MT) is the task of translating one natural language into the other without changing the meaning of the input text and producing a meaningful text in the output language. MT is one of the oldest, hardest, yet extremely useful problem of NLP. NLP is an intersection of fields in Linguistics and Computer Science. It is related to building smart computer systems to understand, analyze and process languages, based on understanding of phenomenon and human perception of that language, to produce desired outcomes like Translation, Sentimental analysis, etc.

Deep Learning (DL) has led to the invention of some great data-driven machines, that can solve extremely complex tasks for humans. These machines comprise of Neural networks, which trains on vast amount of data and are capable of producing desired outcomes. Neural Machine Translation (NMT) is a field of end-to-end learning to specifically focus on automated translations.

Hence, we attempt to build a working model for Neural Machine Translation, a Machine Learning tool which can do automatic language translations. For proof-of-concept purpose, we are going to develop an DL model which can convert German Language to English and vice-versa. In this report, we will present our research as literature survey on the various works that has been done in NMT field. The report also contains our study and understanding of the dataset that we are using to train and test our model. Based on the literature survey and study of available data, we will provide a tentative proposal for solving the problem and discuss various building blocks in the models that we are planning to build.

1. **Literature Survey**
   1. History

Although the history of Language Translation can be traced back to as early as 196 BCE Rosetta Stone [1], a single stone inscribed with three different texts in Ancient Egyptian and Greek languages, we will try to focus on the brief history of Machine Translation, when the idea of computers to do translations started evolving.

The very first idea of “machine” translation is credited to be given by French Engineer George Arstrouni in 1932 is to develop a multilingual mechanical dictionary and a multipurpose “Mechanical Brain” which will be capable of translating languages of different nations into each other.

In the 1950s and 1960s research under Erwin Reifler at the University of Washington formed the word-for-word approach. In 1964 the government sponsors of MT in the United States formed the Automatic Language Processing Advisory Committee (ALPAC) which provided a report in 1966 concluding that MT was slower, less accurate and twice as expensive as human translation and that 'there is no immediate or predictable prospect of useful machine translation.' At Montreal, research began in 1970 on a syntactic transfer system for English-French translation. a group from IBM published in 1988 the results of experiments on a system based purely on statistical methods. What surprised most researchers (particularly those involved in rule-based approaches) was that the results were so acceptable, almost half the phrases translated either matched exactly the translations in the corpus, or expressed the same sense in slightly different words. Works of Hutchins [2] and Poibeau [3] can be referred for detailed historical context of Machine Translation.

Slocum [2] has given a very nice historical perspective of defunct systems in the field of MT. Some to mention are Georgetown’s GAT from 1952 which was developed for translating Russian texts to English under the support of US government but didn’t provide good output quality, Texas’s METAL for German to English MT in 1956 which was funded by US government and Brigham Young University’s ALP system from 1973 which was established to translate Mormon ecclesiastical texts from English into multiple European languages and emphasized on Machine aided translation.

* 1. Challenges

As mentioned in the Problem Definition section, Machine Translation is not an easy task, just because of the sheer amount of ambiguity and diversity in all the languages. Koehn P and Knowles R [4] has described six challenges related to Neural Machine Translation: Lower Adequacy, Insufficient training data, low-frequency words from highly-inflected categories like verbs, long sentences, possible divergence by Attention models for word alignment and Beam search decoding. There is another challenge which is less interpretability of NMT models.

Zhang and Zong [5] have defined that in NMT formulation, sentence is the basic input for modelling. However, some words in the sentence are ambiguous and the sense can only be disambiguated with the context of surrounding sentences or paragraphs. And when translating a document, we need to guarantee the same terms in different sentences lead to the same translation while performing translation sentence by sentence independently cannot achieve this goal. In a word, sentence-level translation will harm the coherence and cohesion of the translated documents if we ignore the discourse connections and relations between sentences. Other common challenges are there may be rare words in the sentence, different language have different syntax.

Apart from generic challenges, there may be special challenges for different language pairs. For example, Use of Kanji in Japanese requires different NMT model architecture, consisting of CNN layers, for the identification of symbols. As our project deals with German-English pair, there are some challenges specific to this pair also. Popovic, Stein and Ney [6] have elaborately discussed about special challenges related to German Compound words to Statistical Machine Translation. Compounding in German means combining several words together to create a new word. Based on simple Permutation, this can increase the vocabulary size drastically. Their paper used a combination of techniques, like splitting and rejoining German compounds and accordingly joining English components as well as performing word alignments based on the splitting point of compound words to improve the translation quality. Koehn [7] proposed the splitting based on Corpus, rather than linguistic, for the same compounding problem.

* 1. Methods/Techniques

There are three main subtypes of Machine Translation- Rule Based, Statistical and Neural. Rule-based is not used anymore now, as it requires specific rule formulations for both the source and target languages based on linguistic information.

Statistical Machine Translation (SMT) [8] doesn’t need any human defined rules, rather it requires vast amount of translated data in both the languages. SMT models attempts to create hypothesis based on the patterns it identifies on analyzing the existing translated data. This hypothesis is then used for fresh translation of new phrases. The ideas of STM were first given by W Weaver in the late 1950s. IBM Research is developed one of the first successful SMT machines in early 1990s that didn’t require any example or rule-based information. Quality deteriorates as sentences gets longer or ambiguity in texts increases as we move away from technical or scientific texts.

Concept of Neural Machine Translation was first introduced by works of Kalchbrenner and Blunsom [9] in 2013. They proposed 2 Recurrent Continuous Translation models- RCTM 1 and RCTM 2 which, based on probability distribution over the source language sentences, determines the target sentence. The two models differ in the sense that first one uses Convolutional Sentence Model (CSM) while second one uses n-gram CSM.

Unlike the Statistical MT, which evaluates on small phrases of sentences for translation and rejoining later, Neural MT creates a vast Neural Network by training on the given corpus, and provides a single target language output for the source language full input sentence. Google switches from phrase-based SMT to GNMT system (Google Neural Machine Translation) [10] in 2016, 10 years later of launching its first translation services in public domain.

Following figure, taken from Google AI publication [11] in 2016, shows the better translation quality of Neural (GNMT) over phrase-based (PBMT) but still underperforms when compared to human translation quality. Score ranges from 0-6, 0 being “completely nonsense translation” and 6 meaning “perfect translation”.

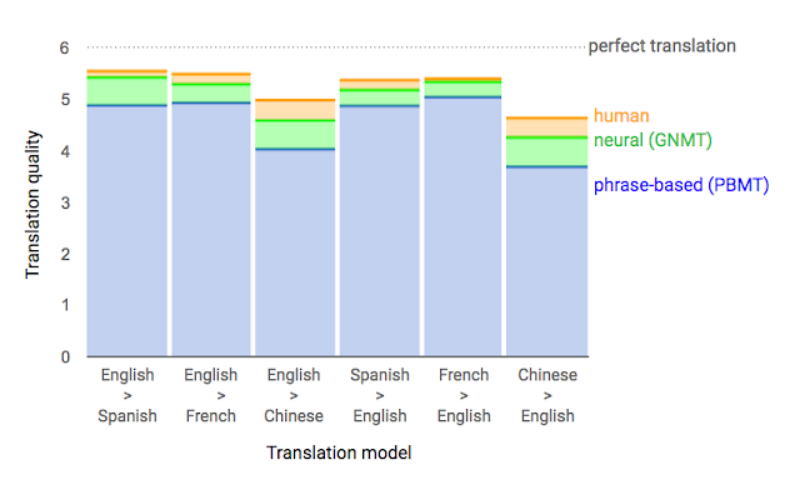


Figure 1. Quality of various NMT models as per Google AI blog.

The use of Attention mechanism for NMT draws its roots from the paper published by Bahadanu et al., 2014 [12]. They introduced an encoder-decoder model which can learn to align and translate jointly and gave very good results for translation of big sentences in single output. Idea was to get rid of fixed length vector to represent input sentence and rather predicting the relevant words at each step and provides than annotation and more weight.

Most of the proposed NMT early were based on a layered architecture of encoder-decoder, created using RNNs (Sutskever et al., 2014 [13]). Later, LSTM based NMT (cho et al., 2014 [14]) was introduced. Vaswani et al., 2017 [15] presented their famous work on Transformers in “Attention is All You Need” publication at NIPS, which got rid of RNNs thus improving the computational speed of MT systems.

* 1. Latest Developments and Comparisons

Google scientists in 2013 released a paper [16] in which they presented their research on similarities among various languages and how to exploit them for Machine Translation. They conducted experiments on WMT11 Datasets for similar language pairs like English and Czech and also on distant language pair of English-Vietnamese. The focus was on improving the dictionaries and phrase tables, which in-turn are used in MT to get good accuracy on short sentence translations. For creating dictionary and representing every word, they used Continuous Bag-of-Words and Skip-gram model [17]. Zhang and Liu [18] has proposed Paragraph-Parallel based NMT with hierarchical attention model to get context from word-level and clause-level abstractions in a structure and dynamic manner.

After Transformers, several new language representation models were created based on the similar technique.

Devlin et al, 2019 [19] introduces BERT (Bidirectional Encoder Representation of Transformers) which improved the accuracy of NMT by masking several words in the source sentence in pre-training phase. This helps overcome the unidirectionality constraint of self-attention, i.e., attending only previous tokens while evaluating current token in the sequence.

Zhu, Jinhua, et al., 2020 [20] conducted many experiments by creating various BERT-fused models for different language pairs. They proposed performing masking into all layers rather than serving on input embeddings only.

Following is a comparison table between difference EN-DE (Engish-German) Translation models that we have discussed so far:

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Model | Dataset | BLEU Score |
| 1 | Phrase Based MT (Freitag et al.,2014) | WMT Test 2014 | 20.7 |
| 2 | RNN Enc-Dec (Luong et al., 2015) | WMT Test 2014 | 11.3 |
| 3 | RNN Enc-Dec Att(Chung el al., 2016) | WMT Test 2015 | 23.45 |
| 4 | GNMT word pieces (Wu et al.,2016) | WMT Test 2014 | 24.61 |
| 5 | Standard Transformer (Vaswani et al., 2017) | ISWLT 2014 | 28.57 |
| 6 | BERT- fused model (Zhu, Jinhua, et al., 2020) | ISWLT 2014 | 30.45 |

Table 1. BLEU score Comparison of various MT models.

Phrase-based MT, although gives considerable performance as compared to very simplistic RNN model, is not very cost effective in terms of time complexity. Introducing Attention in RNN based models improved the score two times, which is massive improvement. GNMT, which is Google translator, has BLEU score of 24.61 when input and output were fed in form of word pieces. Transformer and BERT were proved to be best performing models.

1. **Data**

Three datasets were provided for this use case:

1. Europarl
2. News Commentary
3. Common Crawl

The first set of provided data is mainly taken from version 7 of the [Europarl corpus](https://statmt.org/europarl/), which is freely available. Second set of training data is taken from the new News Commentary corpus. There are about 50 million words of training data per language from the Europarl corpus and 3 million words from the News Commentary corpus. A third resource is the Common Crawl corpus from 2013 which was collected from web sources. Each parallel corpus comes with an annotation file that gives the source of each sentence pair.

For each dataset two text files were provided, one containing German sentences and the other containing the English sentences. The objective was to use the two files to map the German sentences with English sentences.

|  |  |  |
| --- | --- | --- |
| **Name of Dataset** | **No. of sentences in German** | **No. of sentences in English** |
| Common Crawl | 2,399,123 | 2,399,123 |
| Europarl | 1,920,209 | 1,920,209 |
| News Commentary | 202,002 | 201,854 |

*Quality Checks:*

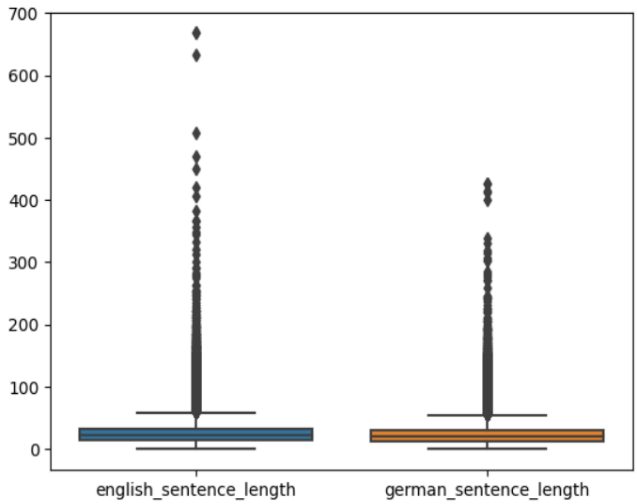
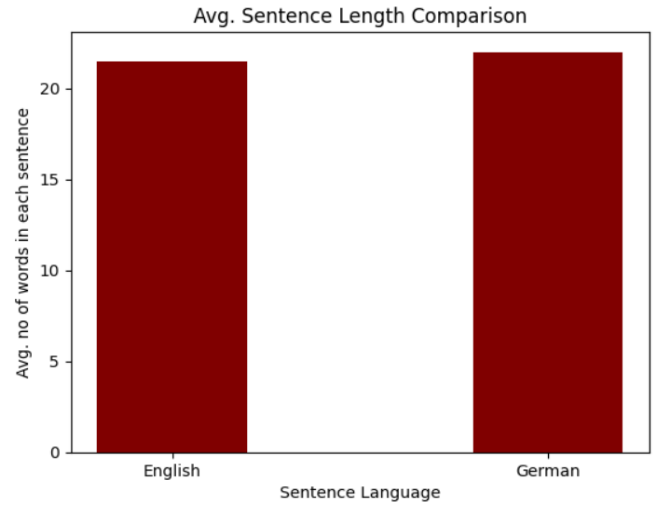
-> Common Crawl dataset contained a few English lines within the German text File

-> Even after trying several preprocessing steps, like removing special characters, removing redundant spaces, removing one letter words, digits and blank spaces, it was not possible to map the two text files

-> For Europarl dataset, the mappings seemed to be consistent. For validation sentences from random indices were selected and put into Google translate. The results returned were satisfactory and it was in-line with the mapping that we were able to do.

Hence, it was concluded that the model training and validation will be done using the Europarl dataset.

For initial insights, the average length of both English sentences and German sentences were compared. Along with that the word distribution in sentences for each language was checked. Below are the results of the plots:



1. **Tentative Roadmap**

From our literature survey, this is what we have understood:

The neural machine translation is an advanced model compared to the phrase-based translation model. In this approach there is no need in building models for each sub-phase of a sentence, rather this approach includes everything within one single large neural network. A decoder then, at that point, yields an interpretation from the encoded vector. The entire encoder-decoder framework, which comprises of the encoder and the decoder for a language pair, is mutually prepared to expand the likelihood of a right interpretation given a source sentence. An expected issue with this encoder-decoder approach is that a neural network should have the option to pack all the vital data of a source sentence into a fixed-length vector. This may make it challenging for the neural network to adapt to long sentences, particularly those that are longer than the sentences in the training corpus. To resolve this issue, the authors acquaint an extended version with the encoder-decoder model which learns to adjust and interpret mutually. Each time the proposed model produces a word in a translation, it soft-searches for a bunch of positions in a source sentence where the most significant data is concentrated. The model then, at that point, predicts an objective word in view of the context vectors related with these original positions and all the past generated target words.

Based on this study, we are planning to implement 3 models:

1. RNN based Encoder-Decoder base model. - 4 layers
2. Bi-directional LSTM based model with Attention mechanism. - 4 layers
3. Transformer based model with Self-Attention mechanism. - 6 layers

Initial number of layer that we are planning to use for Encoder-Decoder in our model to have better sequence-to-sequence handling are mentioned across the model above. Results will be compared for both the models and the best one will be selected for validation.

Recurrent Neural Network (RNN) is a kind of Neural Networks that contains “Hidden” state to capture the sequence related information, thus makes it better at sequence data modelling. LSTM (Long Short-Term Memory) is an advanced kind of RNN, which contains “cell” state that allows it to remember past data in memory easily. This helps in overcome problems of vanishing and exploding gradient which RNN are prone to. Transformers are models that are essentially build on multi-head self-attention mechanisms and contain 6 layers of Encoder and Decoder, but without RNNs.

Transformers are faster than RNN-based models. There are less parameters in transformer model to train. Also, transfer learning is only possible in Transformers. Hence, Transformers are now state-of-the-art building blocks for sequence-to-sequence modelling and hence NMT.

We are going for modular approach. We will develop modules that can be simply imported by user, keeping the complex code abstracted and making it nice and easy to use.

import machine\_translator as mt

conv\_text1 = mt.translate\_eng\_to\_de(“../Data/german.txt”)

conv\_text2 = mt.translate\_de\_to\_eng(“../Data/english.txt”)

print(conv\_text1, conv\_text2)

We will try BERT model as well where masking is performed for the initial embeddings, based on the works of Devlin et al. The expectation is to get a better performance in terms of speed and quality for transformer and BERT model as compared to RNN. Attention mechanism also is expected to improve performance as compared to base model consisting of only RNN.

We are going to use BLEU (Bilingual Evaluation Understudy) [21] score to compare the quality of different models.

1. **References**
2. Andrews, Carol AR. *The Rosetta Stone*. London: British Museum Publications, 1981.
3. Hutchins WJ. Machine translation: A brief history. In Concise history of the language sciences 1995 Jan 1 (pp. 431-445). Pergamon.
4. Poibeau, Thierry. *Machine translation*. MIT Press, 2017.
5. Slocum, Jonathan. "A survey of machine translation: Its history, current status and future prospects." *Computational linguistics* 11.1 (1985): 4-10
6. Zhang J, Zong C. Neural machine translation: Challenges, progress and future. Science China Technological Sciences. 2020 Oct;63(10):2028-50
7. Popović, Maja, Daniel Stein, and Hermann Ney. "Statistical machine translation of German compound words." *Advances in Natural Language Processing: 5th International Conference on NLP, FinTAL 2006 Turku, Finland, August 23-25, 2006 Proceedings*. Springer Berlin Heidelberg, 2006..
8. Koehn, P., Knight, K.: Empirical Methods for Compound Splitting. In: Proc. 10th Conf. of the European Chapter of the Association for Computational Linguistics (EACL), Budapest, Hungary, pp. 347–354 (2003)
9. Brown, Peter F., et al. "A statistical approach to machine translation." *Computational linguistics* 16.2 (1990): 79-85.
10. Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1700–1709. Association for Computational Linguistics.
11. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” (2016).
12. Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team. “A Neural Network for Machine Translation, at Production Scale.” (2016).
13. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
14. Sutskever, I., Vinyals, O., and Le, Q. (2014). “Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems.” (NIPS 2014).
15. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. “Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014). to appear.” (2014).
16. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
17. Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. "Exploiting similarities among languages for machine translation." *arXiv preprint arXiv:1309.4168* (2013)
18. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
19. Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”  *ArXiv* abs/1810.04805 (2019).
20. Zhang Y, Liu G. Paragraph-parallel based neural machine translation model with hierarchical attention. InJournal of Physics: Conference Series 2020 (Vol. 1453, No. 1, p. 012006). IOP Publishing. 2. Cho, K., Van Merri ̈enboer, B., Bahdanau, D., Bengio, Y.: On the proper-ties of neural machine translation: Encoder-decoder approaches. arXiv preprintarXiv:1409.1259 (2014)
21. Zhu, Jinhua, et al. "Incorporating bert into neural machine translation." *arXiv preprint arXiv:2002.06823* (2020).
22. Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.