Akhil Saxena 18BCE0969

Machine Translation Research Review

"Machine Translation: A Literature Review"

Summary:

Machine Translation is the automated process in which one natural language is translated to another natural language. It was first introduced by Petrovich Troyanskii in 1939. Earlier models for machine translation were rule based, but now wit the advent of neural network, lots of research is focused in using neural networks for it so much so that machine learning is almost at par with manual human translation. This paper focuses on Statistical & Neural Machine translation.

Rule-based methods require extensive knowledge of the source language which is to be processed. This becomes complicated for certain languages which have dissimilar structure and sentence/word boundaries. For a statistical method, semantics & syntax alone aren't enough for translation.

For Neural based methods, the most commonly used phrase transition model is the conditional probability of generating a target language phrase given the source language phrase. There is no definition of particular trick of phrase similarity in such models. Most promising method is to use continuous text datasets as the source for training model. This can be then vectorized and neural networks can train on that. Such continuous representations, as opposed to a word-based vocabulary helps capturing the morphological properties along with syntactic and semantic ones.

One important thing is the RNN Encoder-Decoder Architecture which serves as base for most machine translation models. It is a neural model which learns the conditional distribution over variable length sequence.

Some of the methods which have been an area of research in recent times:

1. RCTM: Recurrent Continuous Translation Model is a unique model having 2 components, a generation aspect and a conditional aspect. The generation aspect is handled by RNN and the conditional aspect is modeled by CNN. This is the first task which has based translation purely on the neural network without the use of any statistical system.
2. CSM: Convolutional Sentence Model has a hierarchical structure, similar to parse trees, which enables it to create a sentence representation.

Limitations:

- Conventional Neural Machine Translation systems are not able to handle rare words. These words are called OOV (out-of-vocabulary) words, these are the words which are nonexistent in the source data used for learning.
- Unsupervised methods are yet to reach a quality as good as supervised ones.

"Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation"

Summary:

This paper proposes a single model which can be used to translate between various languages. The model proposed achieves a similar score to what is obtained with separate bilingual translation models. The proposed model does not require changes in the standard architecture of NMT as it only adds a component in the beginning to specify the target language.

The experiments done in the research include mixing all sorts of pairs of languages and translating between them while managing to keep the context intact and fast response times.

All experimental data is checked against the WMT'14 Benchmark and BLEU score. One of the major leaps in this paper is the introduction of Zero-Shot Translation, which basically means translating between a language pair which does not exist in the dataset used to train the model. For example, assume we wish to translate from Hindi to Japanese and there is no data related to conversion of this pair. However, we have Hindi-English and English-Japanese pairs, zero shot translation first translates from Hindi to English in the background and uses that to translate the English text to Japanese.

There are 3 major takeaways from the paper, which are this models best features:

1. Simplicity as adding a new language is just adding more data.
2. Low resource usage as all languages share the same parameters in the model
3. The model can learn to translate between language pairs which do not exist in the dataset.

Limitations:

1. Translation between pair of languages not present requires multiple translations which can hamper the response time as well as cause loss of information or context when searching.
2. Mixing words from different languages causes inability to identify the language and results in the text not being translated to the target language but kept in one of the source text language itself.

"Six Challenges for Neural Machine Translation"

Summary:

Neural Machine Translation has been widely researched in the recent years and has had a lot of progress too with many big companies investing in the research for the same. Compared to Statistical Machine Translation, NMT still faces some issues which are yet to be addressed or are currently being addressed by researchers.

The paper deals with six major challenges for NMT; we summarize each one:

1. *Out of domain*: To identify the context of the text that is present is a difficult task as one word can be used for different purposes when it is used in different contexts.

   For example, dust may mean the physical object itself. (as in, "There is dust on the table"), or it can mean the process of cleaning it. (as in, "Will you dust the room?")

2. *Amount of training data*: Unlike SMT, NMT models do not have a fixed increase when increasing the amount of data in dataset.

3. *Low-frequency words*: Out of vocabulary words still remain a limitation of Machine Translation, be it NMT of SMT.

4. *Long sentences*: Sentences which have a higher word count have been observed to have been translated poorly as compared to when the same sentence was split into multiple fragments. Using attention models, it is remedied a little. However, more research is being done in this area.

5. *Word alignment model*: The alignment of words in a sentence to preserve the true meaning of the sentence is still lagging behind in research.

6. *Beam search decoding*: Though there is a proportionate relationship between the beam size parameter and model score of translations based on it, when the size of the search space is increased, it does not consistently result in better quality of the translations.

Limitations:

1. While NMT models are superior to their Statistical counterparts in various ways, one thing that is not mentioned is how difficult the NMT models are to analyze. It is a monumental task to find the reasoning behind a particular choice an NMT has made in translating from one language to another, whereas in SMTs, we can directly find it.

References:

1. Garg, Ankush & Agarwal, Mayank. (2018). *“Machine Translation: A Literature Review*”.

2. Johnson, Melvin & Schuster, Mike & Le, Quoc & Krikun, Maxim & Wu, Yonghui & Chen, Zhifeng & Thorat, Nikhil & Viégas, Fernanda & Wattenberg, Martin & Corrado, G.s & Hughes, Macduff & Dean, Jeffrey. (2016). “*Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation.”* Transactions of the Association for Computational Linguistics.
   5. 10.1162/tacl_a_00065.

3. Koehn, Philipp & Knowles, Rebecca. (2017). Six Challenges for Neural Machine Translation. 28-39. 10.18653/v1/W17-3204.