CS 410 – Text Information Systems - Tech Review

Student: Maria Fernandez

Neural Machine Translation: History and current trends

Several years ago my manager asked me to look at her son's Spanish homework, knowing I was a native speaker. She wanted to find out if he had used an internet translator instead of writing the Spanish paragraph himself. After reading the text I quickly realized that it was unlikely that a person with basic level of Spanish had written the paragraph. I identified some common shortcoming of automatic translator used at the time. For example, instead of translating "come back" as "Regresar (to return)" it translated it as two separated words "Ir (to go)" and literally "Espalda (back)", the body part.

This story made me realize how far automatic translators have come; therefore I decided to explore the history of machine translation, more importantly the introduction of Neural Machine Translation as a groundbreaking paradigm in this field.

Machine Translation (MT) refers to the process of using computing power to translate a text from a Source Language to a Target Language. One of the first attempts to propose a model for Machine Translation was done by Warren Weaver in 1949 when he was the director of the Natural Sciences Division of the Rockefeller Foundation. He wrote a memorandum to other scientists where he proposed ideas to improve the word-to-word paradigm.

One of the main Machine Translation challenges is ambiguous words. In his memorandum, Weaver proposed to examine the immediate context (words to the left and right) of ambiguous words to help finding the right meaning in the specific text. He also suggested applying cryptographic methods which had proven to be successful during the war. Some of those techniques included putting attention to frequency of letters or letter combinations. He even mentioned the statistical irregularities of languages as well as linguistic universals, which was much later the foundation of statistical machine translation. It is important to notice that these were not only Weavers' ideas, he was also compiling methodologies from other disciplines and applying them to the field of Machine Translation. The relevance of this memorandum was not only the new ideas on machine translation but that it encouraged other scientists to formalize it as a research field.

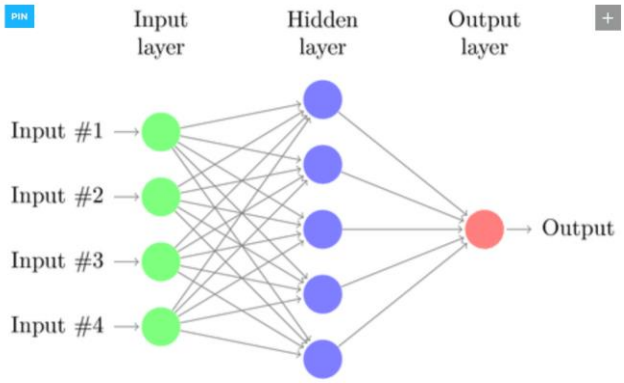
Probably the most notable of Weaver's contributions was to use linguistic information -dictionary and grammars- of the source and target languages to perform MT. This model was later called rule-based machine translation (RBMT) and it remained the main approach in the field until the 1980s when the first Statistical Machine Translation models were developed.

Statistical Machine Translation (SMT) is based in statistical models. The general idea is to start with a large set of approved previous translations that are called corpus which is used to deduce a statistical model of translation. This model is then applied to untranslated texts to suggest a translation. SMT stayed as the main paradigm until the early 2000s when the first neural networks based translation models were developed.

Before continuing the exploration of the uses of Neural Networks on Machine translation, let's define at the most basic level, how neural networks work. Neural networks are a way of machine learning that

allows a computer to learn from analyzing training examples. It is originally model in the basic idea of how the human brain works. It is composed of thousands or sometimes millions of simple interconnected processing nodes organized in layers.

There are several types of neural networks. One of the most common types is called “feed-forward” refers to the idea of the data moving from node to node in only one direction, as shown in this picture:



Picture from https://research.aimultiple.com/how-neural-networks-work/

Each node assigns a weight to each of its inputs and calculates the output by summing up the multiplication of each input value by its weight. Then the node triggers by sending the calculated number to all its connecting nodes in the next layer, if it is above a certain threshold value. The above diagram clearly depicts how the “fast-forward” processing works. The first layer of nodes get an input value which flows to the connected nodes, in that layer those inputs are multiplied and added in complex ways and eventually arrive to the output(s).

A neural network goes through a training phase using a sample data set. During this phase, the weights and thresholds are adjusted until similar inputs yield similar outputs in the set of training data. When the training phase is done, the neural network is ready to be used with unknown inputs and will predict the output with accuracy.

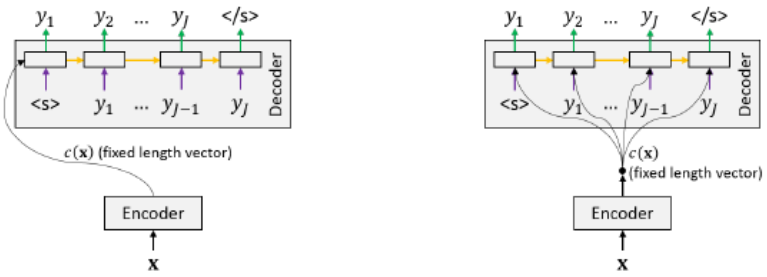
In 2003 Yoshua Bengio and other researchers of the University of Montreal wrote a paper called “A Neural Probabilistic Language Model”. This paper proposed the integration of neural networks into the traditional probabilistic machine translation model to help with issues of sparsity. Sparsity in SMT is caused by the “Curse of Dimensionality” problem. SMT uses vectors to represent terms, the more terms the more dimensions in the vector, the more dimensions the more space between vectors, and that creates the “Curse of Dimensionality” problem, where the space increases so fast that the available data become sparse. Bengio’s solution also addresses the situation when an unknown sequence of words occurs, by that I mean a sequence that does not exist in the training data observed. The main idea is to have a neural network that has a set of parameters that include the vector representations of each word and parameters of the probability function. The purpose of the model is to find the parameters that produce the best prediction of the sample data set. With training, the neural network learns the distributed representations of each word as well as the probability function which is defined in terms of those distributed representations.

Variations and improvements to this model were developed in the next few years, the important thing to notice here is that these models used neural networks as a component of a traditional statistical model versus using neural networks to resolve the complete Machine Translation.

A big shift in the use of neural networks in MT happened fairly recently. Starting in 2013, several researchers were able to develop neural machine translation models that use a single neural network that can be tuned to maximize the translation performance, in other words a neural network that directly transforms a source sentence into the target sentence.

The birth of modern Neural Machine Translation (NMT) is attributed to the work made by Nal Kalchbrenner and Phil Blunsom. They proposed a structure composed by an end-to-end translation model with an encoder and a decoder network, which is still the current prevailing architecture for machine translation. The Encoder Network computes a representation of the source sentence and produces an output in the form of a continuous vector, also referred as state vector. The encoder uses a Convolutional Neural Network (CNN) to do this work. CNN is a type of neural network that is widely used in image recognition and processing. The intuitive idea behind CNNs is that it has a process called convolution that applies a filter in the input which infers interesting features and patters from that dimension. The decoder network, on the other hand, generates the target sentence using a Recurrent Neural Network (RNN) which is a type of neural network that uses loops that allow information to persist throughout different steps in a network. RNNs were used in the decoding process with the intention of addressing the issue of processing text where the current data being translated makes sense based in data previously processed.

The following diagram show two of the original basic architectures of NMT one when the output of the encoder is sent to the decoder once, and the other where it is fed in each step of the decoder.



(a) Source sentence is used to initialize the decoder state.

(b) Source sentence is fed to the decoder at each time step.

Picture from https://arxiv.org/pdf/1912.02047.pdf

As powerful as the new paradigm was, the first implementations had limitations and several improvements to NMT have been made in the recent years. One of them is called Attentional Encoder-Decoder Networks model which improves the quality of translation of long sentences.

The original NMT implementations used a fixed-length vectors for source sentence encoding. This method works great for short sentences but that does not have capacity for long sentences, so NMTs used to chop the source sentences in short clauses. Doing so caused an issue with long distance word re-

orderings which only make sense within the context of a clause. The new model introduces the concept of “Attention”. The idea is that instead of using a fixed-length source sentence representation, it uses a decoder that places its attention only on parts of the source sentence which are useful for producing the next token. In other words, the constant context vector is replaced by a series of context vectors one per each time step. The result of this is that the target word will be predicted based in the context vectors instead of a fix-length vector and it will use weights to put more or less attention them.

Since the birth of NMT, not only the number of academic research efforts has substantially increased, but it has also allowed the creation of publicly available NMT toolkits that are used in production systems by big tech companies. Google, Microsoft, Amazon, IBM among many others, have NMT based toolkits.

Machine Translation has come a long way since Warren Weaver’s memorandum; from linguistic approaches to statistical models to deep learning with neural networks, and it continuously becoming more accurate and efficient. Ambiguity, data context and complex long sentences are still challenges in this field, as in any Natural Language processing activity, but researchers worldwide are working in these problems and finding solutions to them. Some companies like Google have made some of their Natural Language projects open source, like TensorFlow and SyntaxNet, which encourage even further collaboration. Considering how much has been accomplished in the last few years and the attention that academic institutions as well as big tech companies are putting in these projects, leads me to conclude that this field will keep growing and that there are a lot of interesting things yet to come.

Sources:

https://pdfs.semanticscholar.org/eaa9/ccf94b4d129c26faf45a1353ffcbbe9d4fda.pdf

https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

https://research.aimultiple.com/how-neural-networks-work/

https://papers.nips.cc/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf

https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

https://towardsdatascience.com/using-rnns-for-machine-translation-11ddded78ddf

https://ai.googleblog.com/2017/03/an-upgrade-to-syntaxnet-new-models-and.html