One of the earliest goals for computers was the automatic translation of text from one language to another.

Automatic or machine translation is perhaps one of the most challenging artificial intelligence tasks given the fluidity of human language. Classically, rule-based systems were used for this task, which were replaced in the 1990s with statistical methods. More recently, deep neural network models achieve state-of-the-art results in a field that is aptly named neural machine translation.

In this post, you will discover the challenge of machine translation and the effectiveness of neural machine translation models.

After reading this post, you will know:

- Machine translation is challenging given the inherent ambiguity and flexibility of human language.
- Statistical machine translation replaces classical rule-based systems with models that learn to translate from examples.
- Neural machine translation models fit a single model rather than a pipeline of fine-tuned models and currently achieve state-of-the-art results.

Let’s get started.

A Gentle Introduction to Neural Machine Translation
Photo by Fabio Achilli, some rights reserved.

What is Machine Translation?

Machine translation is the task of automatically converting source text in one language to text in another language.

In a machine translation task, the input already consists of a sequence of symbols in some language, and the computer program must convert this into a sequence of symbols in another language.

Given a sequence of text in a source language, there is no one single best translation of that text to another language. This is because of the natural ambiguity and flexibility of human language. This makes the challenge of automatic machine translation difficult, perhaps one of the most difficult in artificial intelligence:

The fact is that accurate translation requires background knowledge in order to resolve ambiguity and establish the content of the sentence.

— Page 21, Artificial Intelligence, A Modern Approach, 3rd Edition, 2009.

Classical machine translation methods often involve rules for converting text in the source language to the target language. The rules are often developed by linguists and may operate at the lexical, syntactic, or semantic level. This focus on rules gives the name to this area of study: Rule-based Machine Translation, or RBMT.

RBMT is characterized with the explicit use and manual creation of linguistically informed rules and representations.

— Page 133, Handbook of Natural Language Processing and Machine Translation, 2011.

The key limitations of the classical machine translation approaches are both the expertise required

What is Statistical Machine Translation?

Statistical machine translation, or SMT for short, is the use of statistical models that learn to translate text from a source language to a target language gives a large corpus of examples.

This task of using a statistical model can be stated formally as follows:

Given a sentence T in the target language, we seek the sentence S from which the translator produced T. We know that our chance of error is minimized by choosing that sentence S that is most probable given T. Thus, we wish to choose S so as to maximize Pr(S|T).

— A Statistical Approach to Machine Translation, 1990.

This formal specification makes the maximizing of the probability of the output sequence given the input sequence of text explicit. It also makes the notion of there being a suite of candidate translations explicit and the need for a search process or decoder to select the one most likely translation from the model's output probability distribution.

Given a text in the source language, what is the most probable translation in the target language? […] how should one construct a statistical model that assigns high probabilities to "good" translations and low probabilities to "bad" translations?

— Page xiii, Syntax-based Statistical Machine Translation, 2017.

The approach is data-driven, requiring only a corpus of examples with both source and target language text. This means linguists are not longer required to specify the rules of translation.

This approach does not need a complex ontology of interlingua concepts, nor does it need handcrafted grammars of the source and target languages, nor a hand-labeled treebank. All it needs is data—sample translations from which a translation model can be learned.

— Page 909, Artificial Intelligence, A Modern Approach, 3rd Edition, 2009.

Quickly, the statistical approach to machine translation outperformed the classical rule-based methods to become the de-facto standard set of techniques.

Since the inception of the field at the end of the 1980s, the most popular models for statistical machine translation […] have been sequence-based. In these models, the basic units of translation are words or sequences of words […] These kinds of models are simple and effective, and they work well for man language pairs

— Syntax-based Statistical Machine Translation, 2017.

The most widely used techniques were phrase-based and focus on translating sub-sequences of the source text piecewise.

Statistical Machine Translation (SMT) has been the dominant translation paradigm for decades. Practical implementations of SMT are generally phrase-based systems (PBMT) which translate sequences of words or phrases where the lengths may differ

— Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

Although effective, statistical machine translation methods suffered from a narrow focus on the phrases being translated, losing the broader nature of the target text. The hard focus on data-driven approaches also meant that methods may have ignored important syntax distinctions known by linguists. Finally, the statistical approaches required careful tuning of each module in the translation pipeline.

What is Neural Machine Translation?

Neural machine translation, or NMT for short, is the use of neural network models to learn a statistical model for machine translation.

The key benefit to the approach is that a single system can be trained directly on source and target text, no longer requiring the pipeline of specialized systems used in statistical machine learning.

Unlike the traditional phrase-based translation system which consists of many small sub-components that are tuned separately, neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation.

— Neural Machine Translation by Jointly Learning to Align and Translate, 2014.

As such, neural machine translation systems are said to be end-to-end systems as only one model is required for the translation.

The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text.

— Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016.

Encoder-Decoder Model

Multilayer Perceptron neural network models can be used for machine translation, although the models are limited by a fixed-length input sequence where the output must be the same length.

These early models have been greatly improved upon recently through the use of recurrent neural networks organized into an encoder-decoder architecture that allow for variable length input and output sequences.

An encoder neural network reads and encodes a source sentence into a fixed-length vector. A decoder then outputs a translation from the encoded vector. The whole encoder–decoder system, which consists of the encoder and the decoder for a language pair, is jointly trained to maximize the probability of a correct translation given a source sentence.

— Neural Machine Translation by Jointly Learning to Align and Translate, 2014.

Key to the encoder-decoder architecture is the ability of the model to encode the source text into an internal fixed-length representation called the context vector. Interestingly, once encoded, different decoding systems could be used, in principle, to translate the context into different languages.

… one model first reads the input sequence and emits a data structure that summarizes the input sequence. We call this summary the "context" C. […] A second mode, usually an RNN, then reads the context C and generates a sentence in the target language.

— Page 461, Deep Learning, 2016.

For more on the Encoder-Decoder recurrent neural network architecture, see the post:

- Encoder-Decoder Long Short-Term Memory Networks

Encoder-Decoders with Attention

Although effective, the Encoder-Decoder architecture has problems with long sequences of text to be translated.

The problem stems from the fixed-length internal representation that must be used to decode each word in the output sequence.

The solution is the use of an attention mechanism that allows the model to learn where to place attention on the input sequence as each word of the output sequence is decoded.

Using a fixed-sized representation to capture all the semantic details of a very long sentence […] is very difficult. […] A more efficient approach, however, is to read the whole sentence or paragraph […], then to produce the translated words one at a time, each time focusing on a different part of he input sentence to gather the semantic details required to produce the next output word.

— Page 462, Deep Learning, 2016.

The encoder-decoder recurrent neural network architecture with attention is currently the state-of-the-art on some benchmark problems for machine translation. And this architecture is used in the heart of the Google Neural Machine Translation system, or GNMT, used in their Google Translate service.
https://translate.google.com

… current state-of-the-art machine translation systems are powered by models that employ attention.

— Page 209, Neural Network Methods in Natural Language Processing, 2017.