М.И.Миняйло
Тюменский государственный университет
Институт математики и компьютерных наук
Кафедра иностранных языков и межкультурной
профессиональной коммуникации
студентка группы КБс-157
manyakos2@yandex.ru
Л.В.Скороходова
Тюменский Государственный Университет
Институт Математики и Компьютерных Наук
Кафедра иностранных языков и межкультурной
профессиональной коммуникации
старший преподаватель кафедры ИЯ и МПКЕНН
l.v.skorokhodova@utmn.ru

ГЛУБОКОЕ ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ ОНЛАЙН-ПЕРЕВОДЧИКА

M.I.Minyaylo
University of Tyumen
Institute of Mathematics and Computer Sciences
Foreign Languages and Intercultural
Professional Communication Department
Student of 157 group
manyakos2@yandex.ru
L.V.Skorokhodova,
University of Tyumen
Institute of Mathematics and Computer Sciences
Foreign Languages and Intercultural
Professional Communication Department
Senior Lecturer

GOOGLE TRANSLATE
THE DEEP-LEARNING UPGRADE OF GOOGLE TRANSLATION METHODS

АННОТАЦИЯ: Существует распространенное мнение, что машинный перевод на базе нейронных сетей вычислительно затратная операция как в тренировке программы, так и в конечном результате перевода. Проблемой является также нарушение устойчивости работы устройства, особенно когда вводимые для перевода предложения содержат редко употребляемые слова, большие объемы информации и длинные речевые модели.

Компания Google представила новую систему для машинного перевода, Google Neural Machine Translation (GNMT), т.е. машинный перевод на базе нейронных сетей. Она использует глубинные нейронные сети для перевода целых предложений, а не только фраз, что значительно улучшает качество перевода.

КЛЮЧЕВЫЕ СЛОВА: глубокое обучение, машинный перевод на базе нейронных сетей, искусственный интеллект, Google Translate, рекуррентные нейронные сети, квантовые вычисления.

ABSTRACT: It seems to be a generally accepted belief that NMT systems are computationally expensive both in training and in translation inference. They also lack of robustness, particularly when input sentences contain rare words, very large data sets and large models.

Google Neural Machine Translation (NMT) has great potential to overcome many of the weaknesses of conventional phrase-based translation systems. Google began experimenting with a deep-learning technique, called neural machine translation that can translate entire sentences without breaking them down into smaller components. That approach eventually reduced the number of Google Translate errors by at least 60 percent on many language pairs in comparison with the older, phrase-based approach.

KEY WORDS: deep-learning technique, Neural Machine Translation, Artificial Intelligence, Google Translate, Recurrent Neural Network, quantized computation.

Probably, each of you dealt with the services of online translators. Their popularity still tends to go up. In this article, we represent the new methods of Google, which allow improving the quality and speed of translation.

Google Neural Machine Translation (GNMT) improves on the quality of translation by applying an example based machine translation method in which the system "learns from millions of examples". GNMT's proposed architecture of system learning was first tested on over a hundred languages supported by Google Translate. With the large end-to-end framework, the system learns over time to create better, more natural translations. GNMT is capable of translating whole sentences at a time, rather than just piece by piece. The GNMT network can undertake interlingual machine translation by encoding the semantics of the sentence, rather than by memorizing phrase-to-phrase translations.

Let us examine the scheme beneath.

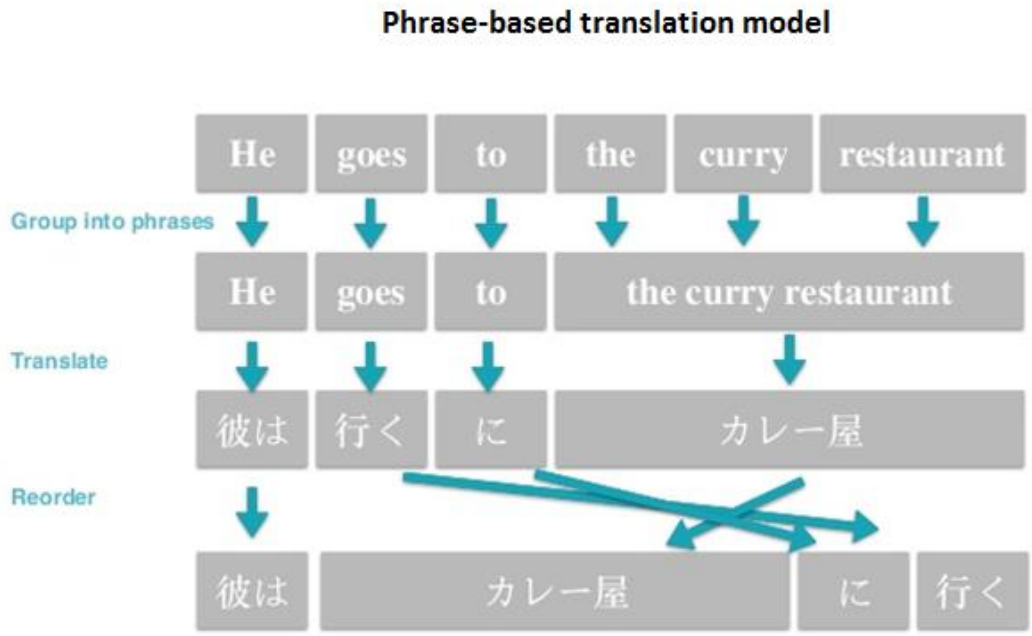

Figure 1. **Phrase-based translation model. The old approach.**

This scheme shows the old “phrase-based” approach that is often used by machine translation services. Here you can see that sentences are broken down into words and phrases to be independently translated. Several years ago, Google began experimenting with a deep-learning technique, called neural machine translation that can translate entire sentences without breaking them down into smaller components. That approach eventually reduced the number of Google Translate errors by at least 60 percent on many language pairs in comparison with the older, phrase-based approach.

The deep-learning approach of Google’s neural machine translation relies on a type of software algorithm known as a recurrent neural network. The neural network consists of nodes, also called artificial neurons, arranged in a stack of layers consisting of 1,024 nodes per layer.

The Model Architecture of GNMT

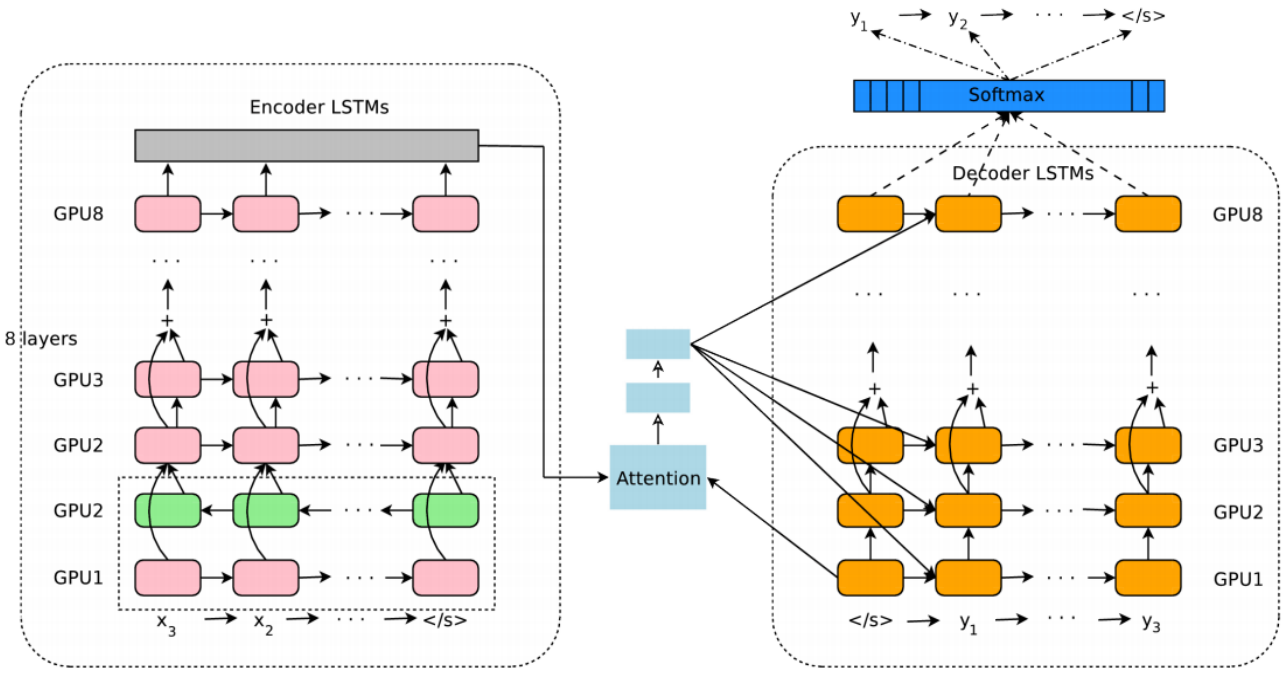

Figure 2. **The Model Architecture of GNMT. The upgrade method.**

Figure 2 demonstrates the model architecture of Google's Neural Machine Translation system. A network of eight layers acts as the "encoder," which takes the sentence targeted for translation—let's say from Chinese to English—and transforms it into a list of "vectors." Each vector in the list represents the meanings of all the words read so far in the sentence, so that a vector farther along the list will include more word meanings. The bottom encoder layer is bi-directional: the pink nodes gather information from left to right while the green nodes gather information from right to left. The other layers of the encoder are uni-directional.

Once the Chinese sentence has been read by the encoder, a network of eight layers acting as the "decoder" generates the English translation one word at a time in a series of steps. The decoder is implemented as a combination of a recurrent neural network (RNN) and a softmax layer. The decoder RNN produces a hidden state for the next symbol to be predicted, which then goes through the softmax layer to generate a probability distribution over candidate output symbols.

A separate "attention network" connects the encoder and decoder by directing the decoder to pay special attention to certain vectors (encoded words) when coming up with the translation. The Chinese sentence splitting into parts, then the neural network selects the appropriate translation, taking into account the weight of each fragment in the text of the original.

This method still generally proved less accurate and required more computational resources than the old approach. Better accuracy often came at the expense of speed, which is also not very cool.

Google researchers had to harness several clever work-around solutions for their deep-learning algorithms to get beyond the existing limitations of neural machine translation. For example, the team connected the attention network to the encoder and decoder networks in a way that sacrificed some accuracy but allowed for faster speed through parallelism—the method of using several processors to run certain parts of the deep-learning algorithm simultaneously.

A third innovation came from using "quantized computation" to reduce the precision of the system's calculations and therefore speed up the translation process. Quantized computation is generally faster than nonquantized because all normally 32-bit or 64-bit data can be compressed into 8 or 16 bits, which reduces the time accessing that data and generally makes it faster to do any computations on it.

Google's neural machine translation also benefits from running on better hardware than traditional CPUs. The tech giant is using a specialized chip designed for deep learning called the Tensor Processing Unit.

When combined with the new algorithm solutions, Google made its neural machine translation more than 30 times faster with almost no loss of translation accuracy.

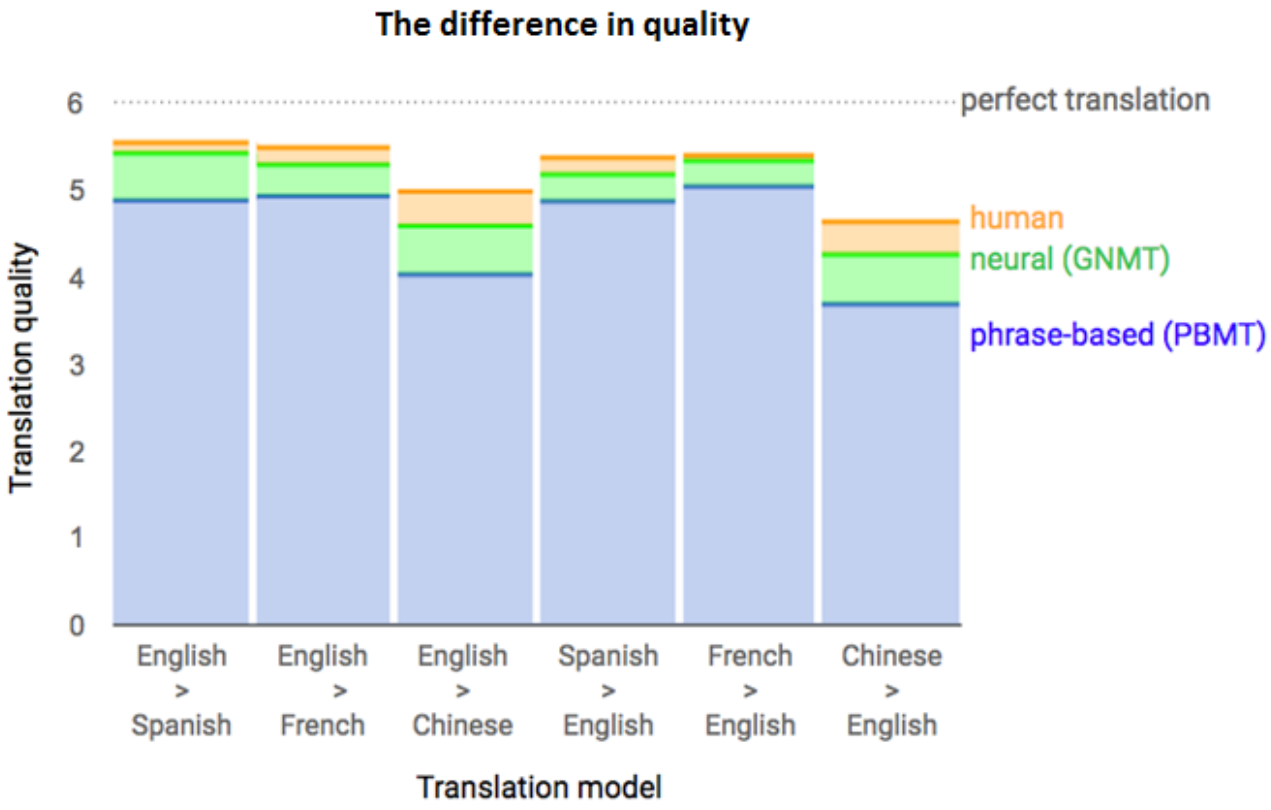

Figure 3. **The difference in quality.**

In addition, on the graph of Figure 3 you can observe the difference in quality of the neural machine translation to the human translation. Moreover, you may notice that the most difficult is the translation of the Chinese-English language pair.

To conclude, Google Translate and other machine translation services still have room for improvement. For example, even the upgraded Google Translate still messes up rare words or simply leaves out certain parts of sentences without translating them. It also still has problems using context to improve its translations. Nevertheless, computer scientists seem optimistic that machine translation services will continue to make future progress and creep ever closer to human capabilities.

References

1. Google Translate Gets a Deep-Learning Upgrade [Электронный ресурс]. URL: http://spectrum.ieee.org/tech-talk/computing/software/google-translate-gets-a-deep-learning-upgrade (дата обращения: 10.11.2017)
2. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units [Электронный ресурс]. 10.06.2016. URL: https://arxiv.org/pdf/1410.8206v4.pdf (дата обращения: 11.11.2017)
3. Wu Y., Schuster M. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation [Электронный ресурс].8.10.2016. URL: https://arxiv.org/pdf/1609.08144v2.pdf (дата обращения: 10.11.2017)