NATURAL LANGUAGE PROCESSING

Google Translate

A System Review: Gaining Insights and Drawing Conclusions

Submitted by Group A - Adriana Kuzmanic, Amala Eggers, Camila Mateo, Frederico Andrade, Luigi Grandi, Marcela Zablah, Sofia Nobre, Usama Khan
Section 2 - IE MBD
Submitted to Professor Angel Castellanos Gonzalez

Index

Introduction

In the following paper, the authors wish to review the Google Translate System, in specific regard to its Natural Language Processing capabilities. Its initiator, Google, has long been on the forefront of digital advancement, providing user friendly solutions to everyday consumer needs, online and free of charge. With this particular system, Google developed a translator that is able to handle a higher difficulty of words, and the meaning beyond singular phrases - two problems with which translation systems have struggled for years. So what makes the current state of the product work so well? How is a machine able to comprehend the multitude of different meanings associated with a lengthy sentence, when altering a word here and there may change its underlying comprehension for a human? According to source 4, Statistical Machine Translation (SMT) has been on the forefront of translation paradigms for the past few decades. Its implementation attempts have been mainly focused on the, priorly stated to be brittle, phrase-based systems. Other projects have combined the language models to learn “phrase representations” with such a SMT system (Source 4), to better relate to meaning. However, these have not yet managed to overcome the trouble of shaking phrase-based approach limitations. Thus, a primary question of the following paper will be how Google Translate manages to outperform previous attempts.

The exploration will focus on both its application functionalities and features, i.e., how a consumer experience runs, and its technical aspects. Herein, the authors will dive into the technologies used to make the system run, the algorithms used and methodologies applied. Further, we wish to portray the current market landscape of the product, such as notable competitors and how they may compare to the system at hand. Before drawing ultimate conclusions about the system, we will aim to classify the product’s nature within a reasonable metric to its competitors, and finally highlight both its future opportunities and limitations. The latter with regard to user experience, technical advancement and market position.

Problem Definition and Choice of System

Before diving into the details of the system at hand, we must assess the business case' relevance in the field and clarify the desired outcome of the exploration. Determining the subject of the review was mainly based on its extreme popularity in the open market. A product so widely used by an array of user groups must hold a certain allure, both technically and from a business perspective. One must consider, of course, the enormous corporation backing the product itself; being integrated into the Google universe, Translate has a unique advantage amongst its competitors. This being exposure or ease of access. We will continue to delve into this aspect of the product in the market landscape section of the paper. We can, however, already establish a high level of relevance within its field, from a public perception perspective alone. Technically, the system is known as highly advanced and holding a complex structure; making it an interesting choice for the purpose of this process.

In order to retain a clear structure to our work, we have defined four guiding topics to be addressed:

- User Functionalities and Features. Here, we will describe the user experience of the system, how it appears and which uses may be applied.
- Technical Review. Diving deeper into the system, this section will cover its functionality, how it is structured and the methodologies employed.
- Market Landscape. As mentioned before, we must first classify how the system fits into its environment amongst other players. A brief competitor analysis will follow, including the state of the art.
- Opportunities and Limitations. This will address where Google translate may excel and where it finds its current shortcomings.

As a result of the above, we will derive our final section, the conclusions and recommendations regarding technical and business aspects.

Usage Review - Functionalities and Features

Since launching in 2006, Translate has become one of Google's most popular assets, serving over 500 million users a month and translating around 140 billion words every day (Source 8). Today, Google Translate manages understanding not only direct translation of direct vocabulary, but also nuances of language and meaning. Extreme advancements can be seen since its initial launch, which seems natural. However, in 2016, Translate managed to outperform consumer expectations and competitor systems. The system was overhauled completely with a new, Artificial Intelligence driven, approach. It is now one of the best consumer-level tools on the open market. The system runs as a stand-alone application, as an integration of Gmail, the browser Google Chrome and a variety of other Google offerings.

The following functionalities can be found on the open platform (Source 9):

- Text translations: Type in text (103 languages).
- Offline translations: Type in text offline (59 languages).
- Instant camera translations: Camera translates in real time (38 languages).
- Photos: Take a photo and upload for more accurate translation (50 languages).
- Conversations: Real-time, two-way conversation (32 languages).
- Handwriting: Draw text and characters on screen (93 languages).
- Phrasebook: Save translated words and phrases.
- Tap to Translate: Tap to copy text in any app and translation pops up (Android only)

The first thing to be considered, when analysing usage of the system, is its multitude of offers. Users are able to not only type in a certain text, but benefit from the variety of convenient alternatives. One in particular - the instant camera translation -, may be handy for travellers, easing their understanding by holding the camera to a menu or road sign, and receive an instant translation.

The User Interface is held in clean white and "google-blue" design, featuring a simple starting page.

The left field serves for text input of the words, or passages, to be translated. The right field then “instantly” displays the text in the selected translation language. The basic, “type in text” function, as stated above, features over 100 foreign (non-english) languages. More complex features, such as the offline function or the camera instant translation function feature 59 and 38 languages, respectively. Another possibility is uploading a photo to the system, where text can be more accurately translated. One feature allows for the translation of real time conversations between two languages, via dictation, while another allows the user to handwrite into the input field the words to be translated. Users can also save certain phrases they wish to retain for future purposes to their system.

Google translate has several other handy features. For example, the system will try to detect most recently accessed languages of your account, in order to set these as the default translation languages, and placing other previously accessed languages at the top of the choice list. This speeds up the translation process, which eases fast communication. When translating a text from English to Spanish, the user can click/ tap a reverse button to switch the direction of the process, while a speaker icon prompts the reading, out loud, of the translated text. Users may find this appealing when having to speak the words they have translated.

Now, while the accuracy of the system has drastically improved and is at the forefront of consumer accessible, free, translation services, it must be stated, that is cannot be seen as a completely reliable source. Businesses or official organizations should rather rely on professional translation services, where a human monitors outcome (Source 10). While the system and artificial intelligence is continuously improving, the technology is not yet infallible.

Technical Review (Algorithm, Technologies, Methodologies)

Behind Google Translate:

In the following section of this paper, we will try to “decode” the reasoning and technical infrastructure of Google Translate. Thus describing, as detailed as possible, the technologies, algorithms and methodologies that allow this tool to be the leader in its own market, as well as the fruitful work of extensive research behind the development of what is, considered a barrier-breaker in the world of linguistics – by allowing us, through translation, reaching different cultures, breaking down walls while travelling, or gaining access to any type of documentation that, otherwise, would be nearly impossible to read, provided we didn't take the time and effort of learning every single language from scratch.

As every other big technology in the market, not every single detail is within our reach, as companies try to keep the secrets behind their innovations from leaving company borders, protecting their R&D from competition. During the introduction of the internet's second phase – Web 2.0 -, translation tools were facing very time-consuming challenges. By using over-simplistic vocabularies with word-to-word translation, difficulties arose: readers had to be mindful of grammar rules, and also make sure every spoken-variation was being accounted for, while translating whole sentences. Currently, these problems have been surpassed. Google Translate can now translate words, sentences and even entire texts just by using the tool in a simple click-output interface. However, how does the Machine Learning translation works?

Machine Learning Approach – The "Google Algorithm":

Almost over 10 years after introducing their tool, Google has created what they have now called GNMT: Google's Neural Machine Translation System, a Google development on Deep Neural Networks. To better understand how GNMT actually works, it's important to delve into certain concepts. One of them being Neural Machine Translation (NMT), which has the potential of addressing many shortcomings of traditional machine translation systems.

NMT reliability focuses on its ability to map from input text to the associated output text - directly -, with a two recurrent neural networks (RNNs) in its architecture: one to consume input text sequence; the other to generate translated output text. NMT is also followed by an attention mechanism which aids the algorithm in coping, as effectively as possible, with long input sequences.

The advantage of using a Neural Machine Translation relies on its ability to overcome design choices usually found in traditional phrase-based machine translation, despite NMT systems relying on worse accuracy – mainly when trained on large-scale datasets, proving an inability to translate less common words. Moreover, NMT systems would sometimes transform certain output sentences, with no similarity, to its attainable translation objective, i.e., it would fail to accurately cover the input, thus resulting in surprising or incorrect translations.

However, Google implemented what they now call GNMT – an advanced NMT system that covers the loopholes of automated translation systems. In this implementation, Google transformed usual RNNs into Long Short-Term Memory (LSTM) RNNs that have, in their core, 8 layers, as well as residual connections between them - to propel gradient flow. As such, Google improved its own system through a set of detailed parameters that would become the steppingstone of Google Translate. These were the following:

- Parallelism – by connecting attention from the bottom layer of the decoder to the top layer of the encoder networks;
- Inference Time – to improve this, Google employs a low-precision arithmetic for inference through special hardware – Google's Wordpieces;
- Use of word pieces (Wordpiece Model) – achieving balance between flexibility of single characters and the efficiency of full words for decoding (deviating from any special treatment to unknown words).

Through this set of parameters, as well as length normalization procedures that efficiently deal with hypotheses comparison problems within different lengths during decoding (inferring a coverage penalty to ensure the model is encouraged to translate all the provided input), Google Translate's model was able to follow a common sequence-to-sequence learning framework, with attention, deeply coded in its own architecture.

Google Translate Model Architecture:

NOTE: please refer to (source 4) "*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016*" research paper, provided by Google, and described in the "References" section of this paper, for a better understanding on the technical details behind Google Translate's implementation systems.

This follows three components: (1) Encoder Network; (2) Decoder Network; (3) Attention Network.

The encoder processes a certain source sentence into a list of vectors (one vector per input symbol). As soon as this list of vectors is generated, the decoder produces one symbol at a time, until it reaches a specifically assigned end-sentence symbol. As such, both encoder and decoder are deeply connected through an attention module that allows the decoder to focus on different sections of a given source section, during the decoding phase.

In Google Translate's architecture, the decoder is implemented as a combination between an RNN network and a softmax layer, i.e., producing a hidden state for the next symbol to be predicted, which then flows to the latter, generating a probability distribution over possible future candidate output symbols.

Through Google's own research, back in 2016, it was discovered that, in order for NMT systems to reveal a good and reliable accuracy, both encoder and decoder RNNs would have to achieve deep enough levels, that would significantly capture subtle irregularities in both source and target languages:

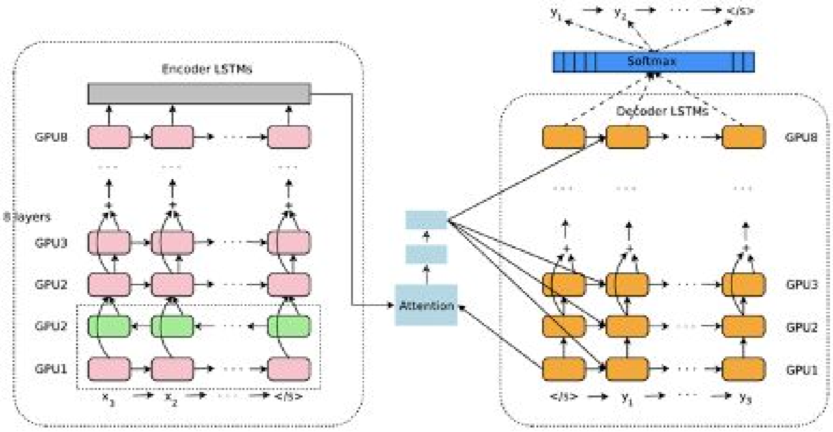


Figure 2 – Google Translate's GNMT Architecture. From left to middle and right: Encoder Network – Attention Module – Decoder Network.

As shown in the graph above, the encoder layer is bi-directional: pink nodes analyze information from left to right, while green nodes analyze it in the opposite direction; and the remaining layers of the encoder are uni-directional. This model is divided into multiple GPUs to achieve a faster training process, allowing, for this current setting, one model replica that is partitioned in 8 ways and placed on 8 distinguished GPUs, each belonging to one host machine. During training, the aforementioned bi-directional encoders compute in parallel, while the uni-directional encoders start computing in separate GPUs. In this architecture, the softmax layer is equally partitioned and placed on multiple GPUs which, depending on the output vocabulary dimension, can either run on the same GPUs as the encoder / decoder networks, or simply run its own separate set of dedicated GPUs.

In addition to its architecture, Google has achieved a deep stacked LSTM – with regards to residual connections –, which predicts better accuracy levels over shallower models. By stacking more LSTM layers, a limit cap on the number of layers is imposed, beyond which the network develops itself, though gradually slower and more difficult to train. As far as Google's research goes, in terms of large-scale translation tasks, a simple stacked LSTM layer will reach peak results within a set of 4 layers, starting to fall with 6 layers, and eventually degrading beyond 8 layers:

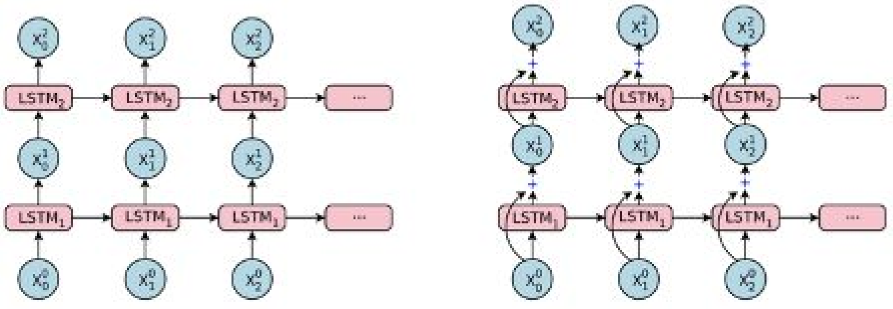


Figure 3 – Normal stacked LSTM (left) VS. Google's stacked LSTM with residual connections (right).

Reclining on the idea of modelling discrepancies amongst intermediate layer-output and its targets, Google introduced residual connections in the LSTM layers, in a stack. Therefore proving that connections will significantly improve the gradient flow, allowing the train of deep encoder and decoder networks.

In this sense, in Google's efforts to implement a deep and well-connected tool such as Google Translate, the information required to translate specific words to an output, can be intertwined anywhere within the source inputted language fed to the tool. In most cases, the source side information is approximately left-to-right, similarly to the target side. However, depending on the pair languages used, the information for a particular output word can be either distributed or splitted in certain regions of the input side. To achieve the best possible context at each point of the encoder network, the bi-directional RNN for the encoder must be used - admitting maximum parallelization during computation. As such, in *Figure 4* (below), it is possible to visualize Google's use of bi-directional LSTMs at the bottom encoder layer: $LSTM_f$ processes the source sentence from left to right, while the layer $LSTM_b$, inversely, processes the same from right to left. Additionally, outputs from $LSTM_f(x_t^f)$ and $LSTM_f(x_t^b)$ are first concatenated, and then fed to the next layer $LSTM_1$.

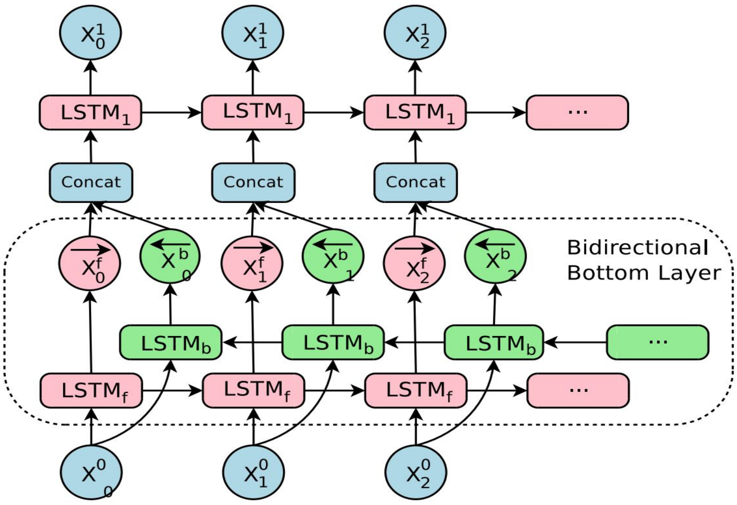


Figure 4: structure of bidirectional connections in the enconder's first layer. $LSTM_f$ *layer processing information from left to right, while* $LSTM_b$ *from right to left. Both outputs are concatenated, and then fed to the next LSTM layer* $LSTM_1$.

Due to Google Translate's architecture model complexity, Google uses both model and data parallelisms to speed up the training process of the algorithms. Data parallelism is straightforward: training a given *n* model replicas, concurrently, using a Downpour SGD algorithm. All the *n* replicas share the same copy of model parameters, with each replica updating the parameters through a combination of *Adam* and *SGD* algorithms. During Google's experiments, they have found out that the number of model replicas should be *n=10*, where each replica works on its own mini-batch of *m* sentence pairs at each time - which again, through Google's analysis, should constitute around *m=128* attempts.

In addition to data parallelism, model parallelism is often used to improve the speed of the gradient computation of each of the replicas. Both encoder and decoder networks are partitioned along each dimension, and placed on multiple GPUs (running each layer on a separate GPU). Hence, the softmax layers are also partitioned, with each single partition accounting for its own subset of symbols within the output vocabulary (as depicted in *Figure 1*).

As for the attention portion of the model, Google chose to align the bottom decoder output to the top encoder output, therefore maximizing parallelism while running the decoder network. Had they aligned the top encoder layer to the top, they would have constrained all parallelism in the decoder network, thus not benefiting from the use of more than one GPU in the decoding process.

Segmentation Approach

NMT models usually operate upon fixed word vocabularies despite translation being, fundamentally, an open vocabulary problem. As such, Google depicts two distinct approaches to be taken, regarding the translation of out-of-vocabulary (OOV) words: (1) one would be to simply copy rare words from source to target - given that rare words consist of name / numbers where a translation would imply a mere copy, either through the attention model, or a more complicated special purpose pointing network; (2) the use of sub-word units - such as characters, mixed word/characters, or more intelligent sub-words.

Given the open vocabulary barriers that Google Translate is subject to, Google has developed two models, towards Segmentation, that are based on the aforementioned approaches:

1. Wordpiece Model:

Google most successful approach (that falls into the category (2) - sub-word units), adopts the Wordpiece Model (WPM) that was intended to initially solve a Japanese / Korean segmentation problem in their first approach within Google Speech Recognition System. This is completely data-driven, guaranteeing a deterministic segmentation for any possible set of characters.

For processing arbitrary words, Google Translate will first break words into wordpieces, given a trained wordpiece model. During decoding phase, the model first produces a wordpiece sequence that is then converted into the corresponding word sequence.

This model is generated using a data-driven approach to maximize the language-model likelihood of the training data, giving an evolving word definition. Google's algorithm to this optimization problem uses a special symbol only at the beginning of the words, not at both ends. Additionally, by removing the number of basic characters to a manageable number (depending on the data) while mapping the rest to a special unknown character - and thus avoiding any disruption to the given wordpiece vocabulary with rare characters -, Google was able to determine that using a total vocabulary of between 8k and 32k wordpieces achieves both good accuracy (BLEU scores), as well as fast decoding speed across all pairs of language that have been tried so far.

Wordpieces achieves, therefore, a balance between flexibility of characters and efficiency of words. Moreover, the models achieve a better overall BLUE when using wordpieces, which derives from the fact that Google's models currently deal with an essentially infinite vocabulary without resorting exclusively to characters.

2. Mixed Word / Character Model

Here, unlike any conventional word model (where usually a fixed-size word vocabulary is kept), the model converts OOV words into the sequence of its constituent characters. Special prefixes are appended to certain characters to not only show the location of characters in a given words, but to also distinguish them from normal in-vocabulary characters.

There are three prefixes: <B>, <M> and <E>, which indicate the beginning, middle and end of a words, respectively. This process is performed on both the source and the target sentences. During decoding, the output may as well contain sequences of special tokens.

Google's Experiments and Results:

Lastly, we will be supporting our research with a given example of Google Translate's methodology of applying GNMT through Google's research paper: *"Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, *2016"* (Reference 4). This research, conducted at Google, provides a benchmark for Neural Machine Translation systems - from English-to-French (WMT En->FR) and English-to-German (WMT En->De), presenting us with the possibility of deep-diving directly into Google's revolutionary methodology. Using these two datasets, the main objective is to depict the contributions of various components of Google's implementation, specifically the wordpiece model, RL model refinement, and model ensembling.

Additionally, we will be able to preview GNMT on Google's translation production corpora. Hence, this study compares the accuracy of this model against human accuracy, and the best Phrase-Based Machine Translation (PBMT) production system for Google Translate.

Datasets

On WMT En->Fr, the training set contained 36M sentence pairs, while the WMT En->De contained only 5M sentence pairs. In both cases, Google used test sets to compare against previous work (a combination of datasets from 2012 to 2014). In addition to the WMTs, Google also evaluated the model on some Google-internal datasets that representer a wider spectrum of languages, which incorporated linguistic properties: English <-> French, English <-> Spanish and English <-> Chinese.

Evaluation Metrics

Google evaluated its models using a standard BLUE score metric. Since the BLUE score cannot quite capture the quality of translation to its fullest, some side-by-side (SxS) evaluations were performed, where human raters evaluate and compare the quality between both translations, for any given source sentence. This SxS score ranges from 0-6, where: 0 =

"Completely Nonsense" and 6 = "Perfect Translation", from meaning to the grammar; and 4 = "sentence retains most of the meaning of the source sentence, but may contain grammar mistakes" and 2 = "sentence preserves some of the meaning of the source sentence but fails in significant parts". These scores were performed by human raters - fluent in both languages -, hence capturing translation quality more accurately than BLUE scores.

Training Procedure

Google trained the models by an already implemented system using TensorFlow, where the training setup follows the classic data parallelism paradigm. With 12 replicas running on separate machines, every single replica updates the shared parameters asynchronously.

All trainable parameters were initialized between the interval [-0.04, 0.04]. As commonly used in RNN model training, a gradient clipping was applied, where all gradients are equally scaled down so that the norm of modified gradients does not exceed the fixed constant (which, in this case, is 5.0).

In the first stage of maximum likelihood training, Google uses a combination of Adam and simple SGD learning algorithms - provided by TensorFlow -, where Adam runs for the first 60k steps, after which a switch is performed to simple SGD. Each step in training contains a mini-batch of 128 examples.

Google found that Adam accelerates training at the beginning, but the algorithm alone converges to a worse payoff, when compared to the combination of Adam-then-SGD:

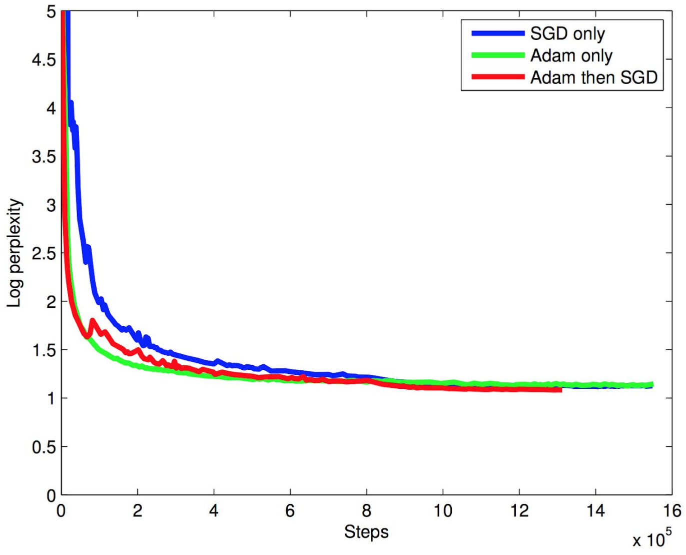


Figure 5 - Log Perplexity VS. steps for Adam, SGD and Adam-then-SGD on WMT En->Fr during maximum likelihood training.

Once a model is fully converged using the ML objective, a switch to RL based model refinement is performed in order to obtain further optimization. The model, therefore, is refined until the BLEU score cannot vary on the development set. Regarding this model refinement phase, Google simply runs the SGD optimization algorithm, given that the number of steps needed for this varies from dataset to dataset - for WMT En->Fr should take around 3 days to complete 400k steps.

Furthermore, to prevent overfitting of the model, Google applied a dropout during training. For the datasets in questions - En->Fr and En->De -, a dropout probability was set to be 0.2 and 0.3, respectively.

Evaluation after Maximum Likelihood Training

The models used in Google's experiment were the following: Word-based, Character-based, Mixed Word-character-based, or Several wordpiece models varying vocabulary sizes.

(1) Word Model: most frequent 212k source words as source vocabulary, as well as the most popular 80k were selected. Words outside of the source vocabulary's scope were converted into special symbols. In addition, the attention mechanism would copy a

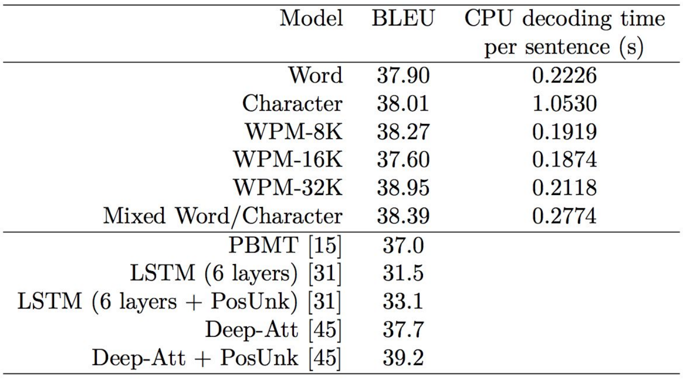
corresponding word from the source, replacing unknown words during decoding phase.

(2) Mixed word-character Model: similar to the above, except the OOV words are converted into sentences of characters with special delimiters. During Google's experiments, the vocabulary size was around 32k. In here, they have simply split all words into constituent characters, which resulted in a few hundred basic characters.

(3) Wordpiece Models: Google trained 3 different models with vocabulary sizes ranging between 8k, 16k and 32k.

The following table shows the results on WMT En->Fr dataset. In here, WPM-32k - wordpiece model with shared source and target vocabulary of 32k wordpieces performs well on the dataset, achieving both best quality and fastest inference speed. As for the pure character model (char input - char output), we can see it performs well on this task, slightly behind the best wordpiece models in BLEU score (however, these models are slower to train and implement, as sentences are much longer).

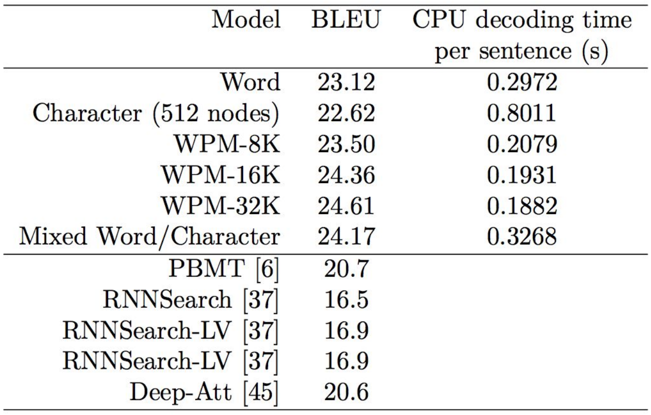
The best model - WPM-32K, achieves a score of 38.95 BLEU (average score of 8 models trained).



Model	BLEU	CPU decoding time per sentence (s)
Word	37.90	0.2226
Character	38.01	1.0530
WPM-8K	38.27	0.1919
WPM-16K	37.60	0.1874
WPM-32K	38.95	0.2118
Mixed Word/Character	38.39	0.2774
PBMT [15]	37.0	
LSTM (6 layers) [31]	31.5	
LSTM (6 layers + PosUnk) [31]	33.1	
Deep-Att [45]	37.7	
Deep-Att + PosUnk [45]	39.2	

Figure 6 - Single model results on WMT En->Fr.

In the same sense, the results of WMT En->De are shown below, where wordpiece models achieve the best BLUE scores:



Model	BLEU	CPU decoding time per sentence (s)
Word	23.12	0.2972
Character (512 nodes)	22.62	0.8011
WPM-8K	23.50	0.2079
WPM-16K	24.36	0.1931
WPM-32K	24.61	0.1882
Mixed Word/Character	24.17	0.3268
PBMT [6]	20.7	
RNNSearch [37]	16.5	
RNNSearch-LV [37]	16.9	
RNNSearch-LV [37]	16.9	
Deep-Att [45]	20.6	

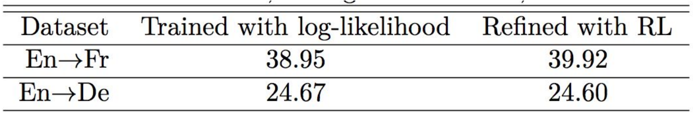
Figure 7 - Single model results on WMT En-De.

In this case, WMT En->De is considered a harder task when compared to the previous dataset, as it contains less training data and German, being a rich language, requires a wider vocabulary for word models. Here, the best model - WPM-32K - achieves a BLUE score of 24.61, over 8 runs. In summary, Wordpiece models tend to perform better than other models both in terms of speed and accuracy.

Evaluation of RL-refined Models:

RL training is used by Google to fine-tune sentence BLEU scores, after the previous maximum-likelihood training.

The results on the best En->Fr and En->De models are shown in the table below, where we can see that fine-tuning models with RL can actually improve BLEU scores: on the first dataset model refinement improves BLUE score by almost 1 point; on the second dataset this improvement is around 0.4; both being the average of 8 independent models.



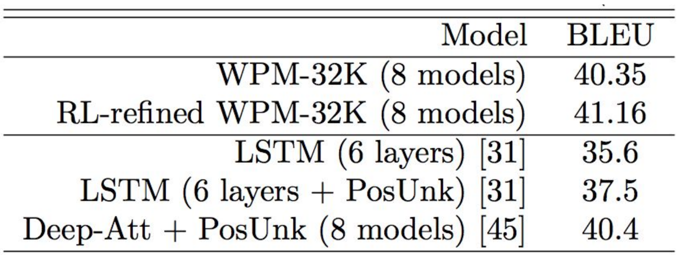
Dataset	Trained with log-likelihood	Refined with RL
En→Fr	38.95	39.92
En→De	24.67	24.60

Figure 8- Single model test BLUE scores, over 8 runs, on both datasets.

Model Ensemble and Human Evaluation

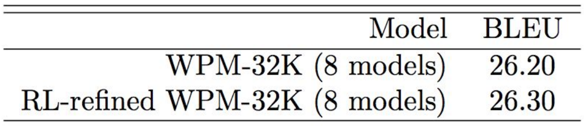
In this evaluation, Google shows the result of 8 RL-refined models on each of the datasets, but then compares the quality of the models (and the RL refinement effect) with a side-by-side human evaluation to observe the NMT translations against reference translations and best phrase-based statistical machine translations - within the foundation of Google Translate.

Therefore, it is important to highlight a result of 41.16 BLEU on the WMT En->Fr dataset (figure 8), and a result of 26.30 BLEU on the WMT En->De (figure 8), both with 8 RL-refined models:



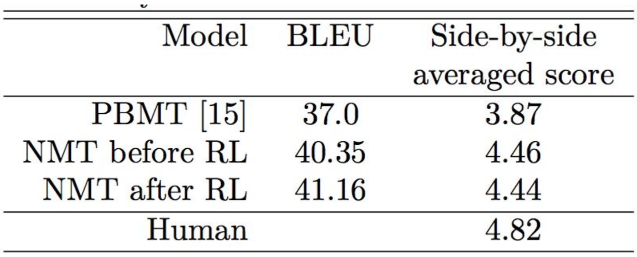
Model	BLEU
WPM-32K (8 models)	40.35
RL-refined WPM-32K (8 models)	41.16
LSTM (6 layers) [31]	35.6
LSTM (6 layers + PosUnk) [31]	37.5
Deep-Att + PosUnk (8 models) [45]	40.4

Figure 9 - Model ensemble results on WMT En->Fr.



Model	BLEU
WPM-32K (8 models)	26.20
RL-refined WPM-32K (8 models)	26.30

Figure 10 - Model ensemble results on WMT En->De.



Model	BLEU	Side-by-side averaged score
PBMT [15]	37.0	3.87
NMT before RL	40.35	4.46
NMT after RL	41.16	4.44
Human		4.82

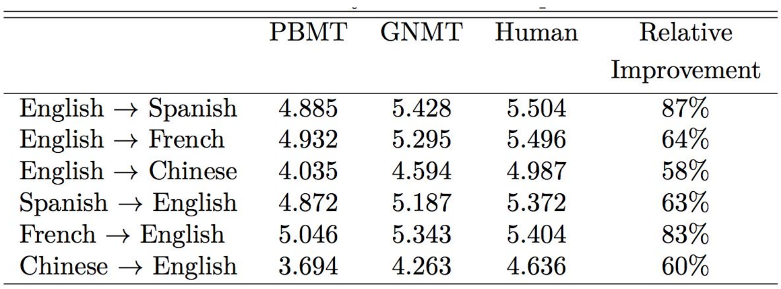
Figure 11 - Human SxS evaluation scores of WMT En->Fr models.

Google found out that even though RL refinement is capable of reaching better BLUE scores, it does not improve human impression of the translation quality. This could be a combination of factors, including: the relatively small size of the experiment; the improvement in BLUE score by RL seems small after model ensembly; the possible mismatch between BLUE as a metric, and real translation quality performed by human raters.

Results of this study:

By asking human raters to rate translations in a three-way SxS comparison, these being: (1) translations from production phrase-based statistical translation system used by Google; (2) translations from GNMT system; (3) translations by humans fluent in both languages; Google was able to carry out extensive experiments on many Google-internal production data sets - as the experiments performed above cast some doubts.

As such, all the GNMT models are wordpiece models, using a shared source and target vocabulary with 32k wordpieces. In the table below, Google shows the average rated scores for English <-> French, English <-> Spanish and English <-> Chinese, given that the evaluation data consists of 500 randomly sampled sentences from Wikipedia and other websites (for each pair of languages), and the corresponding human translation to the targeted language:



	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

Figure 12 - Mean of SxS scores on production data

Results show that the model is able to reduce translation errors by more than 60%, when compared to the PBMT model, on the given set of languages.

Given Google Translate, we can infer that, in some cases, human and GNMT translations are nearly indistinguishable.

As such, by using human-rated SxS comparison as a metric, Google was able to show that the GNMT system approaches the same accuracy as any average bilingual human translator on some of their test sets, delivering roughly 60% reduction in translation errors on several popular language pairs. We can verify that, throughout the course of over 10 years, Google translate is now closer to achieving a more profound and well-rounded interpretation of word combination, thus minimizing language disparities while also giving us a possibility to connect with anyone around - achieving perfect globalization.

Market Landscape - Classification and Competitors

Of course, Google is not the only provider having made advancements in the translation department. Other major corporations, as well as smaller providers, have weighed in on the race toward artificial intelligence powered translation systems. Microsoft Translator, Yandex.Translate, Amazon Translate, IBM Watson Language Translators and DeepL, for example, are relevant competitors. However, not all can be compared to Google Translate directly. In order to understand how Translate compares to its competitors, we have established a system classification process, according to their features. After observing features or the respective competitor's chosen for the classification process, the following metrics stood out as seemingly relevant:

- Translation Approach
- Plattform
- Price
- Language Variety

Google Translate is based on a statistical and neural translation system. It requires no fee and is a cross-platform web application. For the purpose of this paper, we wish to focus on free, all-user accessible platforms, as we believe it is a reasonable overriding category for systems to compare Google Translate to. It has always, and still does to this day, pride itself in aiding humans connect free of charge. Comparing it to more professional and costly systems seems off-topic for now and would eliminate the chance of perfect competition in its viewed market. Thus we can establish:

→ all systems in the following classification are free of charge and of easy access

Let us look further into the translation methods applied as a primary classification index.

Yandex.Translate (Hereafter written as Yandex): Another cross-platform web application, with a SaaS (Software as a Service) license. Yandex, too, applies statistical and neural machine translation. It features 97 languages and was launched in 2011 (Source 11). It is not based on language rules, but retains a more complex, machine translation system.

IBM: The corporate has developed its free commercial version intensively over previous years. Its free version includes a handy feature of document translation while retaining format of the paper, which Google Translate still lacks. However, its number of translation languages is much lower. IBM claims that "By default all language pairs leverage neural machine translation. This new technology uses deep learning to improve translation speed and accuracy" (Source 12). It was originally built on rule-based and statistical models, and is now being improved by the addition of neural machine translation, which however becomes more accessible with premium plans.

Some competitors, such as **NiuTrans**, focus on performing well in a narrow niche-market. This particular system, shows a highly competitive performance for Chinese translation tasks, based mainly on statistical and neural machine translation. It does not, however, manage to deliver to same quality as Google Translate for a broader language variety (Source 13). We therefore choose to rule out NiuTrans and other niche translation systems as being in the same category.

AWS: Based on neural networks trained for language translation, the system is only free up to a certain amount and time of usage. While its performance is strong, it does not show the same accessibility Google Translate can offer (Source 14).

Apertium: A competitor, that is free and a cross-platform application, however translates on a rule-basis, using grammar constraints. As discussed before, this type of translation, close to the system Google Translate used to be, is far from being a accurate in understanding meaning behind phrases. It therefore should not fall into the same category as Google Translate.

From looking at a range of viable competitors, we conclude, that Google Translate should be classified along the following metrics, in a top-performing category:

- Free of charge for the open market
- Cross-platform application with integration capabilities

- Neural Machine Translation Approach for improved accuracy
- Covering high percentage of globally spoken languages

In a traditional market classification, competitors are compared against market attractiveness and market share. Google Translate wins in both aspects. To the average consumer, its name alone stands out for ease of use, effectiveness and accessibility, regardless of it being the most accurate tool in every case.

Opportunities and Limitations of Google Translate

Currently, Google Translate is at the forefront of free translation systems. Both its accuracy and breadth of language coverage, in addition to its many features have set it up for success. One main opportunity for Google Translate in the near future, will be entering the human translator market. One Hour Translation CEO Ofer Shoshan spoke to Forbes Magazine in 2018, claiming that “within one to three years, neural machine technology (NMT) translators will carry out more than 50% of the work handled by the $40 billion market” (Source 15). Shoshan proceeded to state “Today with neural machines, for a growing amount of material and categories, they only need to make a very small number of changes to what a machine outputs, in order to get a human-quality translation”. With a large leap in advancement having occurred, the system is not far off of its end goal, translating to the highest accuracy a human native speaker would be able to. In other aspects, we assume opportunities to lie within the branching out of functionalities and features, such as adapting any competitor’s options. The aforementioned document translation, for example, synonym offerings or automatic dictation language recognition. Further, enabling for API options to non-Google developed platforms, for example instant messaging applications would increase usage and convenience even more, solidifying its position as a market leader. With its extensive financing and branch overreaching

Regarding current limitations, the most obvious one Translate faces is its inability to be used in a professional context. It is simply not at a stage where it can be used for translating contracts or other official documents, and mistakes, especially regarding context or underlying meaning in colloquial language, is sometimes lost on the system. However, reaching back toward our classified category of free systems with a neural machine translation approach, there is no competitor that can reach that goal at the moment. Being the state-of-the-art system, according to our classification, at the moment, also puts a target on Translate’s back. The system is the one to beat, and others are catching up rapidly.

Essentially, innovating itself and remaining in a pole position will be its vital challenge to overcome in the future. From a user experience point of view, Translate would potentially benefit from continuing to improve its intuitive UI (User Interface) even further. For example, a more obvious field for selecting the different features and enabling all features for both Android and iOS.

As far as technicality goes, we can assume that Machine Translation is by no means solved. GNMT will still make significant errors that a human translator wouldn’t. Dropping words and mistranslating proper names or rare terms, in addition to translating sentences in isolation rather than considering the paragraph / page context, still proves that Google Wordpiece has room for improvement. A longer training of the model would be beneficial, in the long run, to Google Translate’s rendition, thus improving the overall score between machine and human perceptions.

Conclusion

The system is no doubt a state of the art translation tool. Its overhaul in 2016 revolutionized the machine translation industry and set the bar for the market high. With its neural machine translation approach, it amazed experts across the globe. How further significant improvements will change the system, and when these may occur, remains hidden to the public. We can be assured, nevertheless, that Google Brain, the top secret think tank of the parent company, established especially for their Artificial Intelligence first initiative, in cooperation with the Google Translate team, will continue to dig deeper into learning processes.

Sources

(1) https://www.kdnuggets.com/2017/09/machine-learning-translation-google-translate-algorithm.html

(2) https://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/

(3) https://en.wikipedia.org/wiki/Google_Neural_Machine_Translation

(4) https://arxiv.org/pdf/1609.08144.pdf (*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, 2016)*

(5) http://www.academia.edu/Documents/in/Google_Translate

(6) https://research.google/pubs/pub45610/

(7) https://www.fit-ift.org/how-does-google-translate-and-similar-tools-actually-work/

(8) https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

(9) https://www.digitaltrends.com/mobile/how-to-use-google-translate-app/

(10) https://www.argotrans.com/blog/accurate-google-translate-2019/

(11) https://www.Yandex.com

(12) https://www.ibm.com/watson/services/language-translator/

(13) https://www.aclweb.org/anthology/W19-5325.pdf

(14) https://docs.aws.amazon.com/translate/latest/dg/how-it-works.html

(15) https://www.forbes.com/sites/bernardmarr/2018/08/24/will-machine-learning-ai-make-human-translators-an-endangered-species/

(16) https://www.digitalistmag.com/digital-economy/2018/07/06/artificial-intelligence-is-changing-translation-industry-but-will-it-work-06178661

(17) https://www.ata-chronicle.online/highlights/artificial-intelligence-and-translation-technology/