
EXPERIENCE
Pure
neural
MT™
SYSTRAN
beyond language

Enter the new era of machine translation

Neural Machine Translation: let's go back to the origins

Each of us have experienced or heard of **deep learning** in day-to-day business applications. What are the fundamentals of this new technology and what new opportunities does it offer?

The concept of deep learning is closely linked to **artificial neural networks**. These networks are already completely changing the way researchers and computer scientists are dealing with complex problems. In the "classic" approach, programmers are giving instructions to the computer by breaking up a problem into a sequence of tasks and algorithms or by relying on big data analysis. In contrast, in an artificial neural network, on one hand, no indications are given to the computer on how to process the problem. Instead, we provide the data that the machine then uses to learn by itself and find solutions on its own. On the other hand, these approaches also contrast with pure statistical approaches based on big data analysis: indeed, neural networks solve problems by generalizing the solution emphasizing the quality of the data, whereas statistical approaches rely on the quantity of data.

Today artificial neural networks and deep learning bring powerful solutions to several domains such as image recognition, voice recognition, digital conversational agents, natural language processing, reasoning and inference applications, ... These solutions have already been deployed on a large scale by Google (Alphago, automatic captioning...), Microsoft (Cortana...) and Facebook.

In the last two years, a lot of research has been conducted on artificial neural networks **applied to natural language processing**. Results are shared among an open source community in which SYSTRAN actively participates.

Impressive results are produced and announced almost daily for applications as various as machine translation, text generation, grammar checking, automatic summarization, chat bots, etc.

What makes NMT a technological breakthrough in the world of machine translation? Jean Senellart, Chief Scientist

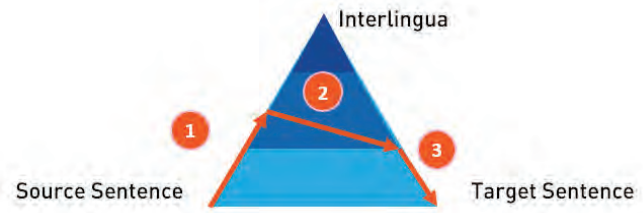
Unlike statistical (SMT) or rule-based (RBMT) engines, **NMT engines process the entire sentence from end-to-end with no intermediate stages between the source and the target sentence**. The NMT engine models the whole process of translation through a unique artificial neural network.
Similar to the human brain, some complementary neural subnetworks combine themselves within this unique neural network while the translation is being generated:
- one subnetwork (or model) addresses the source sentence to extract its meaning,
- a second one, self-specialized in syntactic (grammar) or semantic analysis (word meaning) enriches understanding,
- a third one brings the attention to important words at a given step of the translation
- another one brings contextual knowledge to generate structured and fluent translation,

- finally another model specializes the translation to a domain...

All these subnetworks working together ultimately choose the best translation and **produce a quality that is far better than the current state of the art!**
Neural Machine Translation results are unreasonably impressive, and what is especially remarkable is that in some of the neural network's paraphrasing and mistakes, it seems as if **the neural network truly "understands" the sentence to translate.** In this white paper, we present this technology, how it compares with legacy approaches, the initial results and the expected new applications of this technology.

The automatic translation process

One very simple but still useful representation of any automatic translation process is the following triangle which was introduced by French researcher B. Vauquois in 1968.



The triangle represents the process of transforming the source sentence into the target sentence in 3 different steps.

The left side of the triangle characterizes the source language; the right side the target language. The different levels inside the triangle represent the depth of the analysis of the source sentence, for instance the syntactic or semantic analysis.
We now know that we cannot separate the syntactic and semantic analysis of a given sentence, but still the theory holds that you can dig deeper and deeper into the analysis. The first red arrow represents the analysis of the sentence in the source language. From the actual sentence, which is just a sequence of words, we can build an internal representation corresponding to how deep we can analyze the sentence.
For instance, on one level we can determine the parts of speech of each word (noun, verb, etc.), and on another we can connect words: for instance, which noun phrase is the subject of which verb.

When the analysis is finished, the sentence is "transferred" by a second process into a representation of equal or slightly less depth in the target language. Then, a third process called "generation" generates the actual target sentence from this internal representation, i.e. a meaningful sequence of words in the target language.
The idea of using a triangle is that the deeper you analyze the source language, the simpler the transfer phase is. Ultimately, if we could convert a source language into a universal "interlingua" representation during this analysis, then we would not need to perform any transfer at all – and we would only need an analyzer and generator for each language to translate from any language to any language.

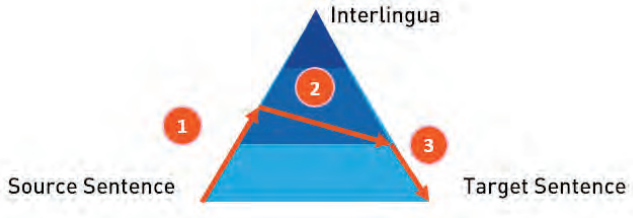
This is the general idea and explains intermediate representation, if any exists, and the mechanisms involved to go from one step to the next. More importantly, mapping and approach to this model forces to describe the necessary resources to accomplish each of the steps.

Let us illustrate how this idea works for the 3 different technologies using a very simple sentence: "The smart mouse plays violin."

Legacy approaches to Machine Translation

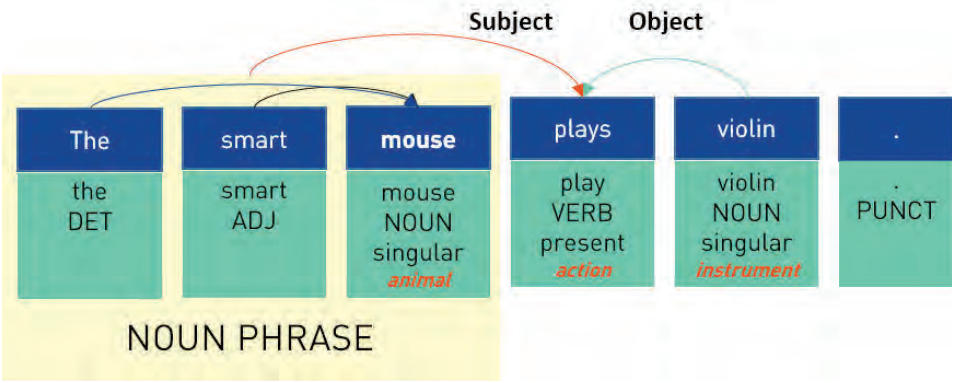
Rule-Based Machine Translation

Rule-Based machine translation is the oldest approach and covers a wide variety of different technologies. This approach is still commonly used for a large variety of applications for its high throughput, deterministic translations critical for some use-cases and its powerful customization ability.



All rule-based engines generally share the following characteristics:
- The process strictly follows the Vauquois triangle and the analysis side is often very advanced, while the generation part is sometimes reduced to the minimal;
- All 3 steps of the process use a database of rules and lexical items on which the rules apply;
- These rules and lexical items are « readable » and can be modified by linguist/ lexicographer.

For instance, the internal representation of our sentence can be the following:

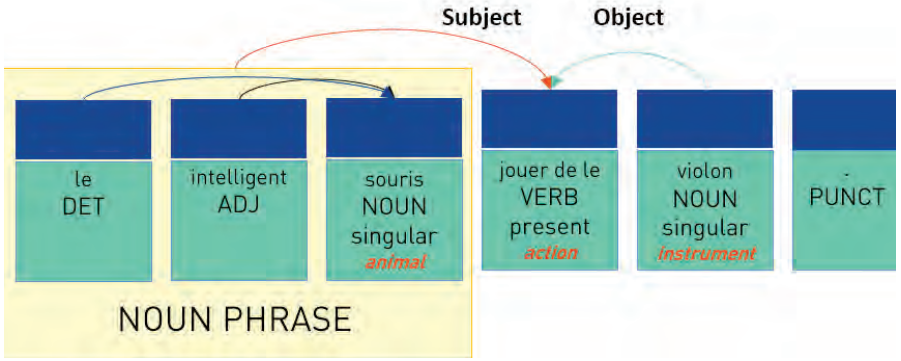


We see different levels of analysis:
- Part of speech tagging: each word is assigned a "part of speech" which is a grammatical category,
- Morphological analysis: "plays" is recognized as inflected third person present form of the verb "play",
- Semantic analysis: some words are assigned a predefined semantic category – for instance "violin" is an instrument,
- Constituent analysis: some words are grouped into a constituent – "the smart mouse" is a noun phrase,
- Dependency analysis: words and phrases are connected with "links," here we identify the subject and the object of the main verb "play."

Transfer of such a structure will use rules and lexical transformations such as:

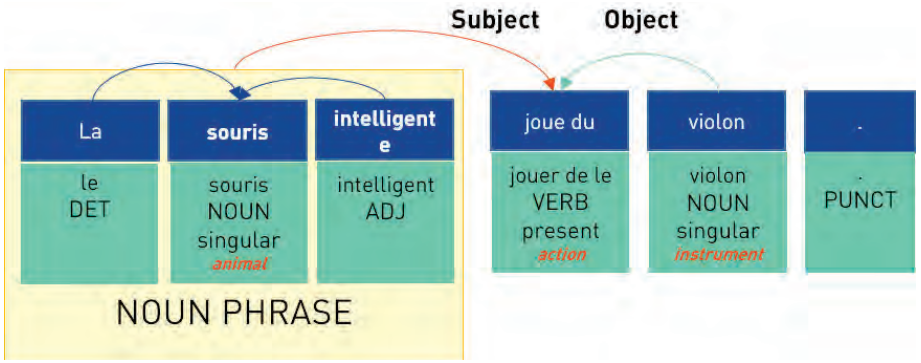
<the.DET>→<le.DET>
<smart.ADJ> modifying <NOUN+animated> → <intelligent.ADJ>
<mouse.NOUN> → <souris.NOUN>
<play.VERB>(subject: S, object: O <NOUN+instrument>) → <jouer de le.VERB>(S,O)
<violin.NOUN> → <violon.NOUN>
<NOUNPHRASE> → <NOUN PHRASE>

Application of these rules on the previous example will generate the target language representation of the sentence:



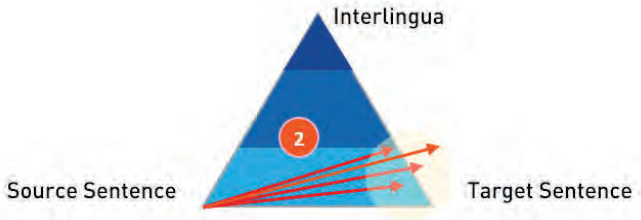
Then French generation rules will be applied, for this example:
- The adjective in a noun phrase follow the nouns – with a few listed exceptions,
- A determiner agrees in number and gender with the noun it modifies,
- An adjective agrees in number and gender with the noun it modifies,
- The verb agrees with the subject...

Ideally, this process would ultimately generate the following translation:



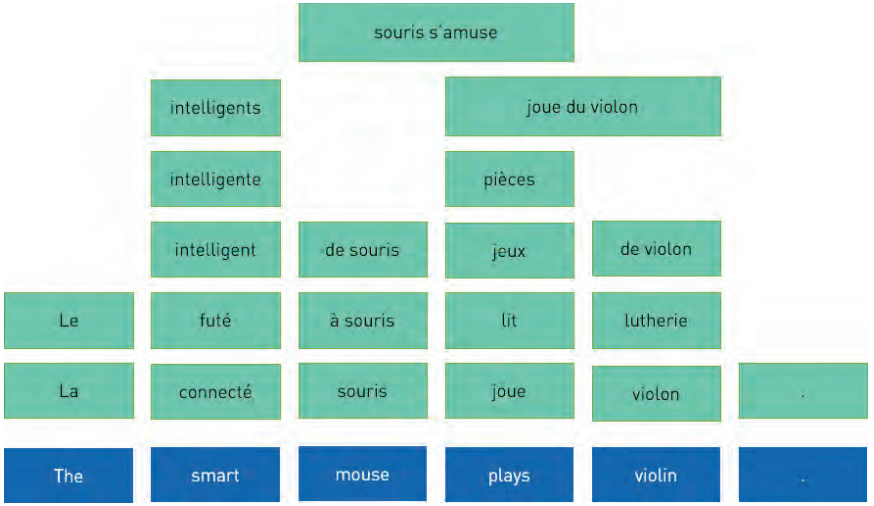
Phrase-Based Machine Translation

Phrase-Based Machine Translation is the simplest and most popular version of statistical machine translation. Till 2016, it was still the main paradigm used behind major online translation services. Its simple training process only requires large volume of existing translation (training corpus) and enables very fast creation of translation model from scratch.
Technically-speaking, phrase-based machine translation does not follow the process defined by Vauquois. Not only is there no analysis or generation, but more importantly the transfer part is not deterministic. This means that the engine can generate multiple hypothetical translations for one source sentence, and the strength of the approach specially resides in its ability to select the best hypothesis.

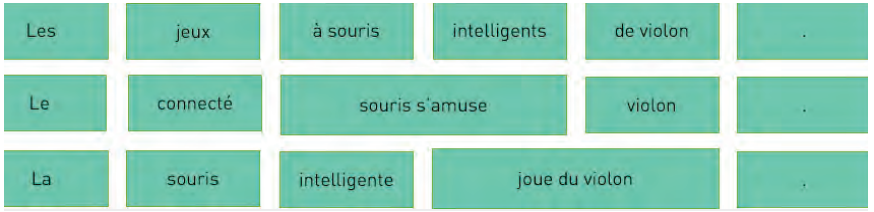


For that, the model is based on 3 main resources:
- A phrase-table which produces translation option and their probabilities for "phrases" (sequences of words) on the source language,
- A reordering table indicating how words can be reordered when transferred from source language to target language,
- A language model which gives fluency to each possible word sequence in the hypothesis.

Hence, from the source sentence, the following chart will be built (in reality, there would be many more options associated to each word):



From this chart, thousands of possible translations for the sentence can be generated, such as the following:



However, thanks to smart probability calculations and smarter search algorithms, only the most likely translations will be explored and ideally the best one kept.
In this approach, the target language model is very important, and we can get an idea of language modeling power simply by doing an online search on the possible options:

Sequence	Online search results = language model search	Comment
Les jeux	34M	This sequence is extremely popular – which means that any translation starting with this sequence will start with a huge advance against alternative translations.
jeux à souris	7	If we have to choose between these two translations, we would take the seconde
jeux de souris	371,000	
souris intelligente	4680	When hesitating between "intelligente" and "intelligent" for translation of "smart" the feminine form will have a slight preference
souris intelligent	4530	

Intuitively, search algorithms prefer to use sequences of words that are probable translations of the source words, with a probable reordering scheme, and generate sequences of words in the target language with a high probability.

In this approach, there is no implicit or explicit linguistic or semantic knowledge. Many variants have been proposed and some show improvements, but to our knowledge and from what we can observe, the main online translation engines use the base mechanism.

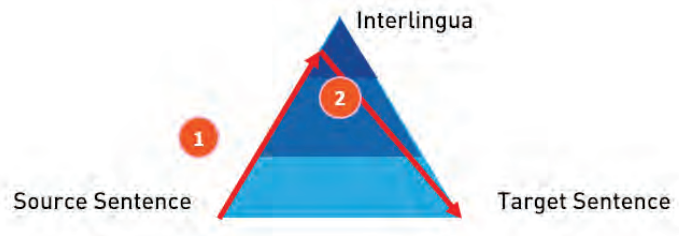
Hybrid Translation

Let us note that state-of-the-art translation providers propose hybrid machine translation engine combining the strengths of rule-based and statistical machine translation. It delivers high translation quality for any domain:

- Rule-based components guarantee predictable and consistent translations, compliance with corporate terminology, out-of-domain usability, and high performance.

- Statistical components learn from existing monolingual and multilingual corpora which drastically reduce customization costs and further improve translation quality within specified domains.

The neural machine translation approach (NMT) is radically different from the previous ones and can be classified using the following Vauquois Triangle:



With the following specificities:

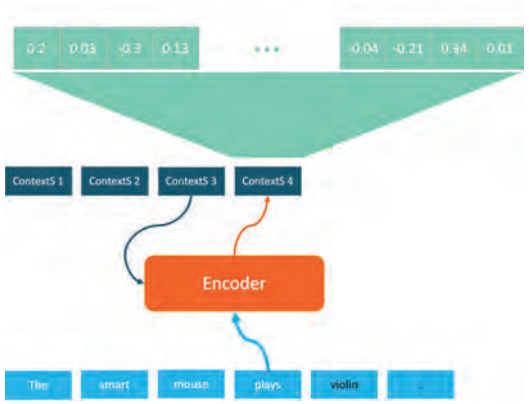
- The "analysis" is called encoding and the result is a matrix composed of sequence of vectors, representing the sentence structure and meaning,

- The "transfer" and "generation" are combined and the process is called decoding and directly generates the target words without any specific generation phase. Note that latest research show that intermediate representation is close to an interlingua representation.

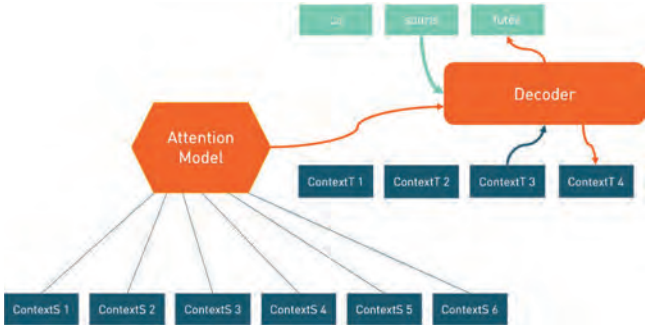
The following sections describe the Global Architecture of a neural translation engine and its building bricks.

Global Architecture

As indicated above, neural translation is based on a 2-step process called encoding/ decoding. Schematically, the "encoder" and "decoders" can be represented as follows:



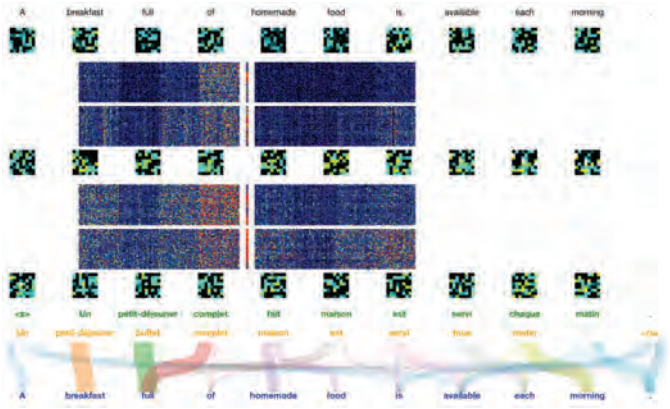
The sequence of source contexts (ContextS 1, … ContextS 5) is the internal representation of the source sentence on the Vauquois triangle and as mentioned above it is a sequence of float numbers (typically 1000 float numbers associated to each source word).



Starting from the sequence of source contexts generated by the encoder, the target words are generated sequentially using the "Target Context" generated together with the previous word, a weighted mix of "Source Context", and the previously translated word using a word embedding to convert the actual word into a vector that the decoder can actually handle.

The translation process ends when the decoder "decides" to generate an end-of-sentence special word. The encoder and decoder are themselves 2 independent artificial neural networks composed of 3 major building bricks described below and often referred by the sentence:

"Embed, Encode, Attend, Predict."



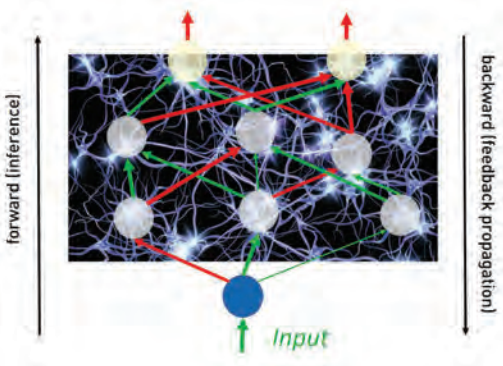
The core technology

Artificial Neural Network and Deep Neural Network

This core technology is remotely inspired by human brain neural network:

- An Artificial Neural Network (ANN) is composed of layers of artificial neurons, the layers are connected together with weights called the parameters of the network.
- Each artificial neuron is activated through simultaneous firing of connected neurons so that an input "signal" is transformed into an output signal through "forward propagation."
- A key element of neural network is in its ability to automatically correct its parameters during the training phase. Technically, the generated output is compared to expected reference and corrective feedback sent "backward" to adjust weights and tune the network connections.

An artificial neural network with more than 3 layers is called a **deep neural network**. Typically, neural networks used for image recognition can have hundreds of layers while neural networks used for natural language processing have 8-20 layers.



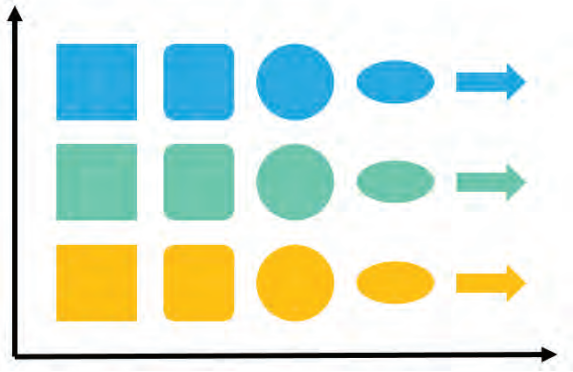
The power of Word Embeddings

Deep Neural Network have been initially successfully deployed for character recognition. For this task, the input of the neural network is an image, which is a sequence of numeric numbers representing each pixel value. For Neural Machine Translation, the "input" is a sentence or a document – and the first step of the process is therefore to convert words into numeric values that the neural network can manipulate.

This conversion is called "word embedding" and the first step of the encoder is to look up each source word in a word embedding table. Part of how meanings are represented in neural machine translation are in the word embedding.

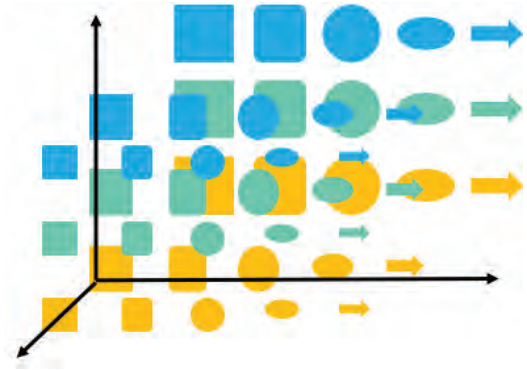
To understand word embeddings, imagine you have to represent different objects with a variation of shapes and colors on a two-dimensional space where objects that are placed nearest to each other should be the most similar.

Below is one possibility:



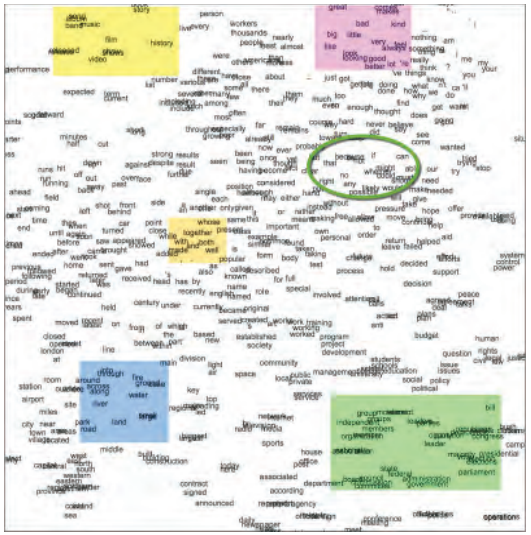
The horizontal axis represents the shape, and we try to place the shapes that are most similar to each other (we would need to specify what makes a shape similar, but for this example, this seems intuitively satisfactory).

The vertical axis represents the color – green being between yellow and blue. If we had different sizes, we could add a third dimension as follows:



If we add more colors or shapes, we might also add more dimensions so that any given point can represent different objects and distances between two objects that reflect their similarity.
The underlying concept in this simple problem is the same than for word embeddings. Instead of objects, there are words, the space is far bigger – generally, we use 300 to 1000 dimensions – but the idea is that words can be represented in such a space with the same properties.

Hence, words with some common property will be near on one dimension of this space. For instance, we can imagine that the part of speech of words are one dimension, their gender if any, another, the fact that they are negative or positive words another, and so on.



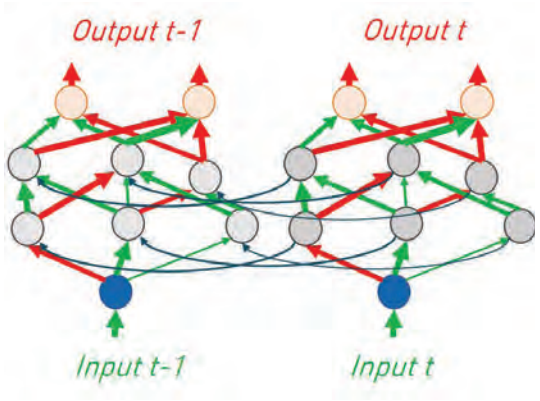
In this graph, a 50-dimension word embedding is projected in 2D. We can see interesting clusters of words:
- in yellow, words semantically related to story, movies...
- in green, political terminology
- in pink, adjectives qualifying goodness/badness
- in the green circle, function words
- in the small orange square, linking words (and, with, both, together...)
- in blue, natural elements (water, fire, ground...).

This embedding has been learned during training process.

The Unreasonable Effectiveness of Recurrent Neural Networks

Deep Neural Networks (DNN) can handle many kinds of data. But originally, they can't use context nor store any knowledge. By default, Artificial Neural Network are indeed "stateless" - which means that the same input given to the network in different context will produce the same output.

For natural language processing, context is extremely important. Indeed, within each sentence, the meaning of a word is very dependent on the rest of the sentence, and even paragraph/documents. Allowing contextual knowledge can be achieved by adding the notion of recurrence in DNN.



In a recurrent neural network, the output of the DNN at a specific time t depends on the input of the timestep t but also on the state of the hidden layers of the timestep t-1.

DNNs, which use recurrence, are called Recurrent Neural Network (or RNN) models and are able to memorize knowledge and sequencing.
There are special kinds of RNNs, called Long-Short-Term Memory (LSTM) RNN, which can mimic human memory by allowing information to be stored in a short or long term term period, several mechanisms compete at each time step to reinforce specific bit of information or to forget it.

Although very simple, RNN demonstrates an "unreasonable effectiveness" to model language at different levels (local agreement, global consistency, fluency, etc...) and key contributors in a NMT engine where generated text is close to flawless and at least generally outperforms non native speaker.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Example of text generated by a RNN trained on Shakespeare poetry. The model captures the style, the document structure, the vocabulary and generates realistic and fluent "Shakespeare sentences".

Attention Model

Another example of DNN mimicking humans is the notion of attention. For instance, when we look at a picture, our gaze focuses on specific parts of the picture where our brain thinks the information is and allows most of the background details to be ignored.
Attention Models for DNN use the same approach: DNN focuses on specific parts of a picture in computer vision, or on specific parts (words) of the sentence in machine translation.

The colored part show where the computer looks to classify the image it is the visual attention.

In much the same way, as humans first focus on important words in the sentence when translating it, the translation attention model focuses on words which it considers most important for the NMT model at a given stage of the translation.
This attention model is essentially an elegant way of keeping track of a memory of DNN source states and only referring to relevant ones when needed.

North	Korean	nuclear	weapons	was	the	biggest	issue	in	the	last	general	elections	.	
0.13328	0.14377	0.15685	0.12976	0.06219	0.00894	0.03101	0.03119	0.00453	0.01970	0.06193	0.06504	0.08136	0.07046	➡ 북한의
0.04826	0.04896	0.30722	0.27316	0.07577	0.00752	0.03498	0.03946	0.01510	0.00821	0.02858	0.03848	0.04288	0.03143	➡ 핵무기
0.04287	0.03131	0.09637	0.33576	0.11143	0.02502	0.07347	0.08131	0.05826	0.03031	0.02760	0.02211	0.02375	0.04043	➡ 문제는
0.03883	0.03364	0.03177	0.03522	0.02886	0.02594	0.05948	0.05953	0.01424	0.05279	0.12439	0.16958	0.14887	0.17684	➡ 지난
0.00684	0.00995	0.01156	0.01135	0.01801	0.00908	0.00827	0.01800	0.01193	0.01904	0.01392	0.22976	0.31111	0.32118	➡ 총선에서
0.02902	0.03327	0.02769	0.03740	0.03683	0.06291	0.17468	0.19540	0.21666	0.03650	0.02090	0.05080	0.03005	0.04791	➡ 가장
0.02130	0.01910	0.01418	0.02022	0.03933	0.04433	0.25615	0.30135	0.16490	0.02601	0.03169	0.02011	0.01274	0.02860	➡ 큰
0.01283	0.02497	0.02325	0.03461	0.04459	0.05369	0.03292	0.46746	0.19363	0.01866	0.00906	0.02374	0.02034	0.04026	➡ 이슈@@
0.01462	0.02029	0.00998	0.02977	0.15925	0.27001	0.03160	0.19532	0.11702	0.04121	0.00376	0.00569	0.00981	0.09167	➡ 였다
0.05679	0.04907	0.03203	0.06615	0.19261	0.11992	0.02477	0.04258	0.02872	0.07112	0.01760	0.01123	0.02180	0.26561	➡ .
0.06220	0.06569	0.06255	0.08600	0.17459	0.06834	0.02992	0.07469	0.02453	0.05396	0.01776	0.02175	0.03172	0.22629	➡ </s>

During translation of the English sentence in the top line, each line is representing a step of the decoding process leading to the generation of the word in the right column.
We visualize the attention of the DNN on the source words with the color range: in green, we see the words receiving the bigger attention, in other words where the model is looking at to generate the target language word.

Human evaluation

First results

A>B	HUMAN	PNMT	BING	GOOGLE	NAVER
HUMAN		54	71.7	62.4	65.8
PNMT	38.5		61.9	60.4	60
BING	18.9	25.8		20.8	26.8
GOOGLE	30.2	26.3	56.2		48.8
NAVER	25.4	28.2	45.9	28	

SYSTRAN PNMT is judged better by human evaluators*:

- than reference human translation in 38.5% of the cases
- than Bing translation in 61.9% of the cases
- than Google Translation in 60.4%
- than Naver translation in 60% of the cases

** Evaluation run mid-October 2016 on English>Korean language pair: the other evaluated systems are PBMT based.*

SYSTRAN PNMT ™ first results

While PNMT is outperforming previous MT generation engine for almost all configurations, language pairs and automatic score, it is interesting to look at output from all the generations to understand strength and weaknesses of each.
Let us consider the verb "to render" - we can find the following meanings in a reference paper dictionary:

to render (service, assistance, ...) to someone ➔ apporter
provide or give (service, assistance, ...)

to render something (adjective) ➔ rendre
cause to be or become (adjective)

to render (a verdict, judgement) ➔ rendre
to deliver (verdict, judgement)

Translation of this some sentences including the different usage of this entry are given in the following table. In this table, the three systems have been trained and tuned on the same training corpus.

SOURCE	PNMT	RBMT	PBMT
The rain renders his escape extremely dangerous.	La pluie rend son évasion extrêmement dangereuse.	La pluie rend son évasion extrêmement dangereuse.	La pluie rend son évasion extrêmement **dangereux.**
Mrs Evans would render assistance to those in need.	Mme Evans **aiderait** ceux qui en ont besoin.	Mme Evans **fournirait l'aide** à ceux dans le besoin.	Mme Evans **prêter** assistance aux personnes dans le besoin.
Mrs Evans would render all her help to those in need.	Mme Evans apporterait toute son aide à ceux qui en ont besoin.	Mme Evans **fournirait toute son aide** à ceux dans le besoin.	Mme Evans **rendrait toute son aide** à ceux dans le besoin.
Mrs Evans would render her verdict to those in need.	Mme Evans rendrait son verdict **à ceux qui en ont besoin.**	Mme Evans rendrait son verdict à ceux dans le besoin.	Mme Evans rendrait son verdict à ceux dans le besoin.

On these simple sentences, we can see the underlying strengths and weaknesses of each of the technology:

- Phrase-Based system is missing long-distance agreement, real contextual analysis and generalization of corpus observations even though it has an incredible memorization ability that the other approaches are missing.

- Rule-Based MT can easily generate non fluent translation while it has the ability to deal with complex rules input by human

- NMT in general is able to learn and deal correctly with contextual rules, has also the ability to paraphrase some sentences - which in some cases can be a weakness. Also, NMT is generating extremely fluent sentences, which is a strength in comparison with human translation: in our tests PNMT generally outperforms non native target language translator.

" From everything I understand from my conversations, SYSTRAN is far along the NMT path, and miles ahead in terms of actually having something to show and sell, relative to any other MT vendor "

Kirti Vashee, Independent Technology & Marketing Consultant

"As far as we are aware, it is the first time access to neural machine translation for a variety of language combinations is offered to the general public [...] In term of quality, there is an actual improvement in fluency, even to the casual tester."

Florian Faes, Slator

From Generic to NMT Specialization

For many of us, we often have to adapt to a new situation, such as a new task. For instance, each time we start a new job, we must adapt to the new situation, a new environment, a new company, etc. Translation tasks share exactly similar issues: adapting to a new domain (from news to legal, for instance), a new style (literature, patent, manuals...), etc. This adaptation is production critical and over the past few years, and is a core feature of SYSTRAN's technologies and services.

For Neural Machine Translation, these approaches can be processed at three levels: before the training, during the training and after the training.

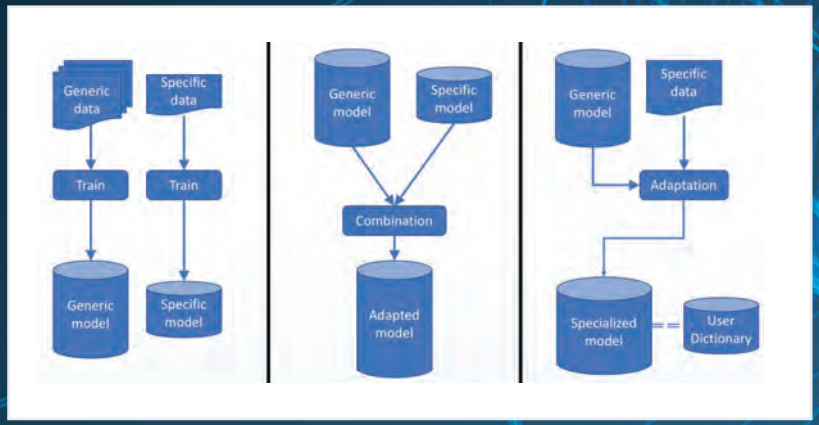
The classical approach proposes learning a model for each specific domain. The corpora used to learn the model are composed of what is called "in-domain" data, which belong only to the domain we seek to translate. However, as this leads to small models, not every common word can be translated and the number of untranslated words named "Out-of-Vocabulary" words (OOVs) increases.
One way to avoid OOVs is the combination of a generic and specific model. There are several ways to combine models, but the results is often a bigger model, which could be difficult to manipulate. Indeed, a bigger model, implies a bigger machine to host the model.

Inspired by Computer Vision (CV) approaches of adaptation, we propose a way to optimize neural networks in a post-training process, which we call "specialization."
Specialization consists of taking a generic model and adapting it to new data without fully retraining it. For instance, let us take our generic model trained on generic data over a period of several days (or even weeks) and relaunch the training process only on a small amount of in-domain data.
The in-domain size can be a single document of around four thousand lines, which is very small compared to the four million sentences of the generic training corpus.

Next steps

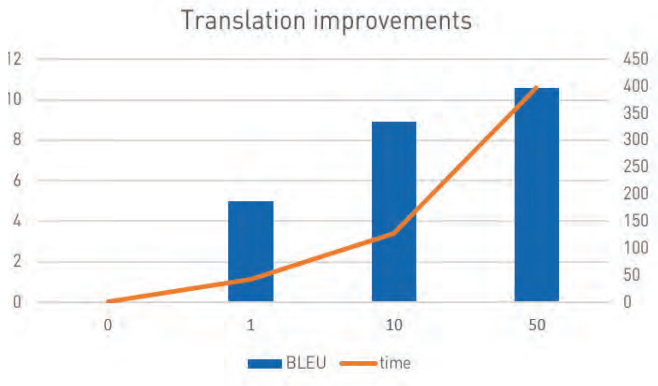
Neural machine specialization

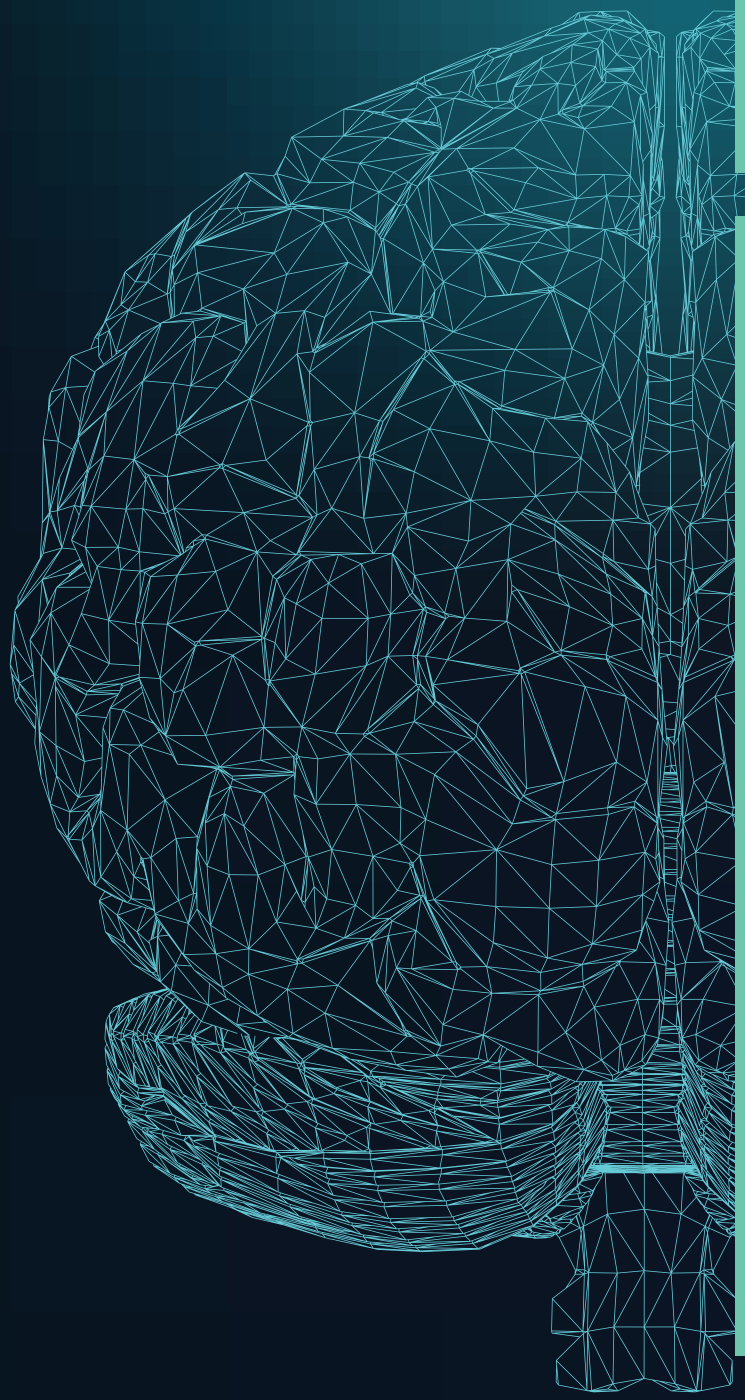
Adaptation Models



This capacity to re-arrange the deep neural network or re-estimate its weight though another training phase is well known in the Machine Learning community.

In the same process, a new specific terminology can be added in PNMT, which guarantees a full domain adaptation.
Using this application in PNMT results in many improvements, as shown in the graphic beside.
By using only one document, it takes less than 50 seconds to adapt the generic model and improve the translation performance by 5 points.
This process is a win-win approach in terms of processing time and performance.
Finally, the specialized model is not greedier than the original model in terms of technical restrictions.

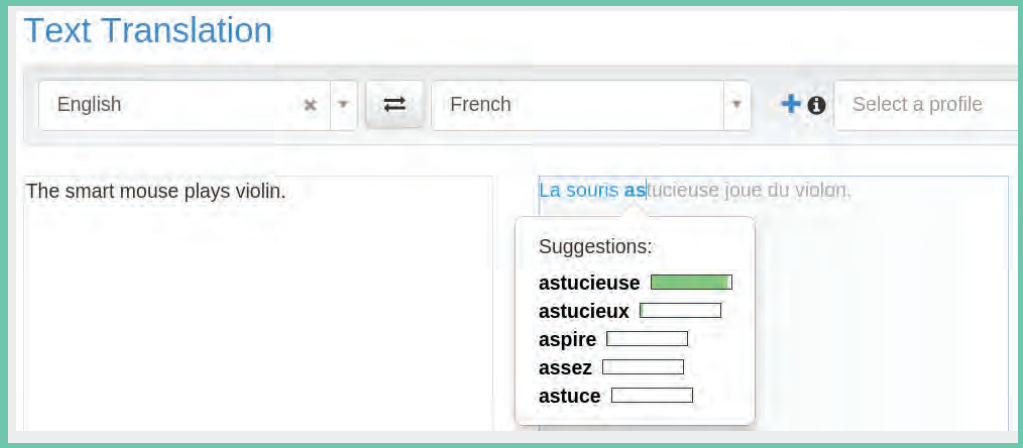




NMT new and future applications

While Neural Machine Translation is undeniably bringing huge improvements in translation quality and will continue to raise the bar, what is also very exciting is the new range of applications that this new technology will enable.

Typically, it is already possible to introduce a full interaction between the human translator and the neural network translation with predictive translation interfaces:



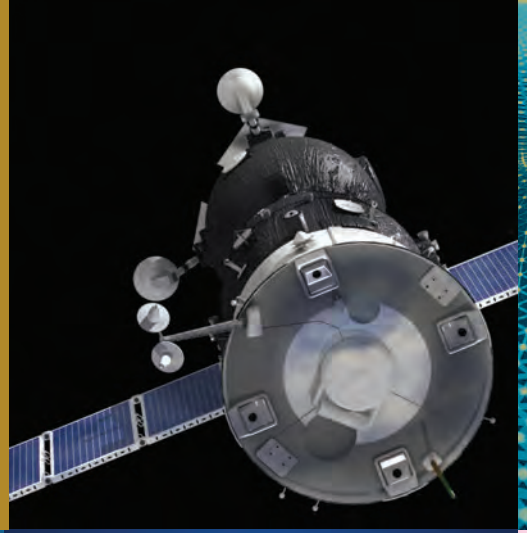
Predictive translation interface: at each moment the system proposes the best completion to the human translator

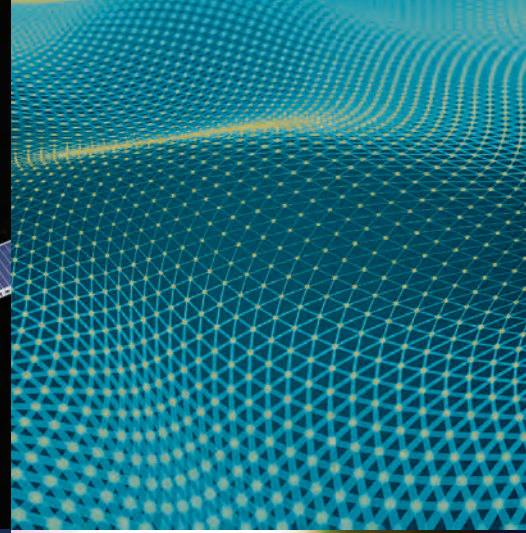
The field of possible applications is endless and our teams are working hard to extend our customer applications in many fields, including:

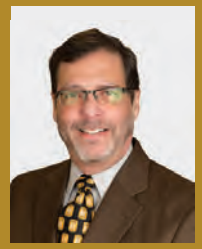
- **translation checking application**
- **multilingual authoring**
- **crosslingual summarization**
- **language learning assistant**
- **multilingual video conferencing ...**

"After a five-year lull in the MT development world and seemingly little to no progress, we finally have some excitement in the world of machine translation and NMT is still quite nascent. It will only get better and smarter."

Kirti Vashee, Independent Technology & Marketing Consultant

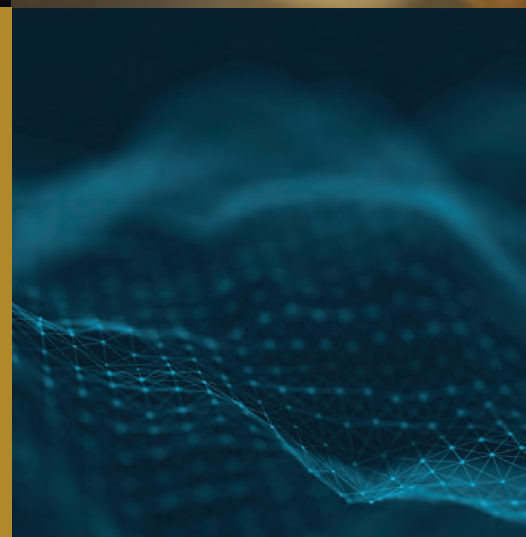






"While today's NMT may be slow and resource-intensive, CSA Research sees this as only the start of the development curve. Software optimizations and improvements in techniques should eliminate this performance penalty over time, as it has in other computation- and memory-intensive applications."

Don A. de Palma, Common Sense Advisory



SYSTRAN

beyond language

We are Systran, we love languages