Department of Translation, Interpretation and East Asian Studies

Master in Tradumatics: Translation Technologies

Research Paper

Corpus-Based Machine Translation: A Study Case for the e-Government of Costa Rica

Gloriana Cocozza Garro

Contact information:

gcocozza.com/glori.cocozza@gmail.com

Director: Gökhan Dogru

Universidad Autónoma de Barcelona

Spain, 2018

Abstract

This research paper aims to approach the state-of-the-art technologies in machine translation. Following an overview of the architecture and mechanisms underpinning PB-SMT and NMT systems, we will focus on a specific use-case that would attest the translator's agency at maximizing the cutting-edge potential of these technologies, particularly the PB-SMT's capacity. The use-case urges the translator to dig out of his/her toolbox the best practices possible to improve the translation output text by means of data preparation, training, assessment and refinement tasks.

Keywords: phrased-based, neural, machine translation, corpora, training, KantanMT, postediting.

Resumen

Esta investigación pretende estudiar el estado del arte en las tecnologías de la traducción automática. Se explorará la teoría fundamental de los sistemas estadísticos basados en frases (PB-SMT) y neuronales (NMT): su arquitectura y funcionamiento. Luego, nos concentraremos en un caso de estudio que pondrá a prueba la capacidad del traductor para aprovechar al máximo el potencial de estas tecnologías. Este caso de estudio incita al traductor a poner en práctica todos sus conocimientos y habilidades profesionales para llevar a cabo la preparación de datos, entrenamiento, evaluación y ajuste de los motores.

Palabras clave: traducción automática, neuronal, basada en frases, corpora, entrenamiento, KantanMT, postedición

Resum

Aquesta investigació pretén estudiar l’estat de l’art en les tecnologies de la traducció automàtica. S’explorarà la teoria fonamental dels sistemes estadístics basats en frases (PB-SMT) i neuronals (NMT): la seva arquitectura i funcionament. A continuació, ens enfocarem en un cas d’estudi que posarà a prova la capacitat del traductor per aprofitar al màxim el potencial d’aquestes tecnologies. Aquest cas d’estudi incita el traductor a posar en pràctica tots els seus coneixements i habilitats professionals per dur a terme la preparació de dades, entrenament, avaluació i ajustament dels motors.

Paraules clau: traducció automàtica, neuronal, basada en frases, corpus, entrenament, KantanMT, post-edició

Contents

Tables

Figures

1. Introduction

Background:

The idea of using machine translation (MT) as a complementary tool to improve the translation process has existed since its creation. With the exception of a few attempts to replace human translators (Fully automatic high-quality translations[1]) during the first generation of MT, today's state-of-the-art-technologies bring new levels of agency for translators to embrace and exponentially benefit from these technologies.

Regarding the private and public sectors and their growing interest in implementing MT systems, a variety of applications have been developed since the 1980s. For instance, Systran was one of the first companies to offer customized MTs to clients such as Aérospatiale, Dornier, NATO and General Motors (as cited in Dolz, p.9). Similarly, public institutions have adopted these systems to cope with the increasing demand of content availability in all EU official languages. On November 15th, 2017, the European Commission (EC), for instance, began to offer a custom machine translation service (eTranslation) for EU institutions and related public institutions. This technological asset would not have been possible without past EU financial funds in research and development. In the ambition of creating a digital and multilingual EU platform, EUROTRA I in 1982, EUROTRA II in 1990, Euromatrix 2006-2009 had aimed to create a machine capable of translating texts related to information technology and EC documents.

Thanks to these solidarity funds, progress in machine translation has been possible. In fact, translation memories represent an indirect outcome of years of research related to this subject, and today they are an essential in our toolbox. Even though many of us now take them for granted, the literature suggests that fuzzy matches are our favorites because they simply boost up translators' productivity. Regarding the impact of MTs to our profession, eTranslation states in its site,

[1] The acronym FAHQT was first coined in 1950 by Yeshoua Bar-Hillel, as noted by Hutchins in "The Practical Use of MT Systems."

> "Our machine translation service produces raw automatic translations. Use it to grasp the gist of a text or as the starting point for a human-quality translation. If you need a perfectly accurate, high-quality translation, the text still needs to be revised by a skilled professional translator." (Machine Translation for Public Administrations,)

Contrary to early misconceptions that placed MT as a threat to human translators, this statement depicts today's landscape, in which translators and MTs are bound to ally because first, to achieve human-like translation quality, translation systems cannot go without human intervention; second, translators cannot overlook the fact that MTs are a feasible and cost-effective solution; third, the latest MT technologies are data-driven, and since we (translators) are the main producers of these resources, this then suggests that translators are the starting point of these technologies.

Among today's ongoing projects, the application of MT on e-governments caught my attention the most, particularly *El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)*, funded by the Spanish government. Its goal is to "support the development of fields such as the natural language processing, machine translation and speech recognition systems in Spanish and co-official languages" (Agenda Digital para España).[2] It consists of four phases. The first phase refers to the process of collecting the most amount of linguistic resources possible (open data in the form of parallel corpora and dictionaries), stored by the public administration and other institutions. The second phase is related to the process of internationalization, which means adapting the content to facilitate its localization. The third phase relates to the process of creating a centralized memory and train a neural MT. Finally, phase four seeks to apply this prototype to specific domains related to the sectors of healthcare, tourism and education.

The particular case of the Plan TL served as a source of inspiration for this research. Just like the government of Spain, Costa Rica (CR) has enforced a digital agenda. In 2014, CR

[2] This citation was originally retrieved in Spanish as "fomentar el desarrollo del procesamiento del lenguaje natural, la traducción automática y los sistemas conversacionales en lengua española y lenguas cooficiales."

accomplished to jump from rank 77 to 54 in the UN program report named "E-government the future that we want"[3] (Salas, L.). The journalist, Salas, further reports that Costa Rica has classified among the top 10 countries in America with the highest improvement on e-government matters. Nonetheless, the fact that the translation industry in CR is not as contemporary as the European one makes evident the need for overcoming an ICT (Information and Communication Technologies) lag of more than 10 years. Based on this context, this research aims to study the possibility to carry out phase three and four of the Plan TL; i.e., to explore the most suitable approach to MT if implemented in the translation of Costa Rica's e-government's content. Perhaps, with the implementation of an MT system technology, we could aspire to develop a bilingual MT similar to Portage, in Canada[4].

To achieve this, we will focus mostly on the tasks of a translator such as data collection, preparation, refinement and assessment of a Phrase Based and a Neural MT. The data collection and preparation and refinement phases will be addressed in the Methodology. In the Literature of the Review, we will learn about the basic components and processes that underpin PB-SMT and NMT systems. Then, in the Analysis and Results chapter, we will assess our machines output text and compare their performance in terms of BLEU, TER and F-measure metrics and human-based evaluations. Finally, the Conclusions will provide a series of final thoughts regarding the experiment and the possible future steps for further research.

[3] The program report was originally retrieved in Spanish as "E-gobierno para el futuro que queremos."

[4] Canada enforced the Official Languages Act in 1972. In 2012, the government spent $2.4 billion in translation services. Then, Portage, a SMT system, was developed to reduce translation costs (Machine Translation and Governments, 2016).

2. Objectives

The primary goal of this project is to build an in-domain machine translation engine. By pursuing this goal, we would be able to determine which corpus-based technology is the most proficient to approach the translation (localization) process of specific content available today in the Costa Rican e-government. Such system must fulfill the following criteria: The engine should be capable of producing an acceptable level of fluency and adequacy. Second, the chosen platform should feature learning capabilities to reduce the post-editing cost. Third, the chosen platform should be user-friendly enough for translators to avail this technology themselves in their work.

Regarding the specific objectives, this paper will attempt:

1. To assess and compare the translation output from an in-domain PB-SMT and NMT.
2. To identify the linguistic challenges, deficiencies and improvements of each of the output texts based on a human-based evaluation.
3. To identify the task-oriented strengths and limitations of each system, based on the automated metrics.
4. To determine which translation output text is more proficient for production purposes.

3. Literature of the Review

3.1 A Brief History of Machine Translation

"If a machine behaves as intelligently as a human being, then it is as intelligent as a human being" Alan Turing, 1950.

Machine translation or 'mechanical translation' is indeed an old technological ambition. In the late 1940s, during World War II, ideas about a computer being able to process natural language began to emerge[5]. Pioneers from all backgrounds held great expectations and fed machines with linguistic knowledge. This led way to the development of the first Rule-Based Machine Translation systems. These machines, however, were very expensive, in terms of training, maintenance and adaptation (Haddow, 2014). In spite of that setback, Way and Hearne point out that, still today, leading companies, like Systran, "offer RBMT engines for purchase" (On the Role, 2011). Another existing approach is the Phrase-Based Statistical Machine Translation (PB-SMT). These engines are the most widely used today, and they work with statistical models. Even though PB-SMTs are not linguistically competent, they render more adequate and fluent translations than RBMTs in specific use-cases. This is due to the use of bilingual and monolingual corpora, which are huge collections of natural language.

Another important approach to MT is the encoder-decoder-attention model. This technology applies artificial intelligence and deep learning techniques[6]. According to the Farajian, A. et al. (2017), although PB-SMT has vastly dominated research recently, now more attention is driven towards NMTs, particularly because they have accomplished to overcome some of the limitations of PB-SMTs. As an example, Marcello points out that these systems, particularly ModernMT, allow internal training and "on-the-fly adaptation",

[5] Weaver had first mentioned the possibility of using computers to translate in March 1947, in a letter to the cyberneticist Norbert Wiener" (Hutchins, 2000).

[6] Alternative implementations of this technology nowadays are speech and handwriting recognition systems such as Cortana and Siri. As cited in Chung, J.

which means a great reduction in pre-production time (the engine adapts to the domain instantly), and ultimately in cost.

Although these new features of NMTs represent the beginning of a shift towards a new era of MT, the other technologies, RBMT and PB-SMT, show similar to better performance in contrast to many cases of NMT systems. Their performance depends on many factors and scenarios. Just like Marcello affirms, “NMTs are superior only on some languages.” In other use-cases, the fact that NMT’s translations are so natural sounding might lead to counterproductive post-editing costs. Other factors that affect the performance of a system are the language combinations, the quality of the data and the levels of quality expected. As a result, in light of achieving the research objectives and in acknowledging the two most dominant technologies in the industry, the following section aims to explain the state of the art approaches (PB-SMT and NMT), their architecture and major components.

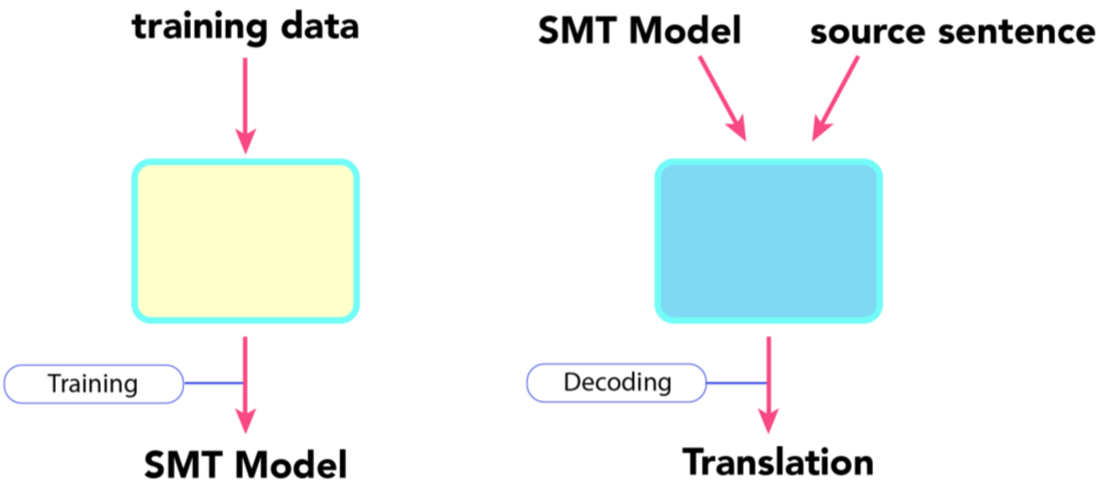
3.2 Phrase-Based Statistical Machine Translation

As mentioned before, Phrase-Based Statistical Machine Translation (PB-SMT) is a data-driven approach. In other words, it “simply learns to translate from already existing human translations” (Kenny and Doherty, p.278). Examples of well-known PB-SMT systems nowadays are IBM, NiuTrans, Google Translate[7], Bing Translator, Yandex, among others[8]. In addition, as its name suggests it, a data-driven engine applies a number of statistical models or ‘features’ to search for the most probable translation. As seen in Figure 1, a PB-SMT follows two distinct phases: training and decoding.

[7] Google Translate used mostly PB-SMT technologies until September 27th, 2016, when it announced the launch of a Google Neural Machine Translation (GNMT).

[8] For a more complete list, visit Comparison of machine translation applications. Today, Google Translate and Microsoft Translator Hub make use of neural systems.

Figure 1. The Training and Decoding Phases of a PB-SMT



The training phase consists of creating a statistical system (SMT Model) based on the training data. Apart from that, SMTs work under the influence of many different models, but the most relevant ones for the scope of this research are the Translation Model and the Language Model. Following Kenny and Doherty's explanation, the TM is trained on the bitexts and glossaries; and the LM is trained only on monolingual corpora of the target language. Now, to understand the purpose of each of these models, Kenny et al. explain,

> "so rather than just ask whether 'the house' is a likely translation of la maison (the answer to which question should come from the translation model), the SMT system also needs to ask whether 'the house' is a likely sequence in English in the first place."

Based on this statement, translation models give more weight to adequacy; whereas, language models care more about fluency.[9] To clarify this and before explaining in detail the components that form a PB-SMT Model, it is important to at least acknowledge the algorithm behind the mechanics of a SMT Model. This model can be expressed by a single algorithm named the log-linear model. Given an input source sentence, this formula aims

[9] A mid-phase between training and decoding in PB-SMT is called tuning. During this phase, the user can readjust the weights of the models. Most commercial platforms feature auto tuning and auto testing options, which means that the engine would automatically take representative segments from the training data to tune and test the engine's performance.

to find the most likely translation of that sentence. Figure 2 shows the formula of probability:

Figure 2. The Log-linear Algorithm. Taken from Perez, M.

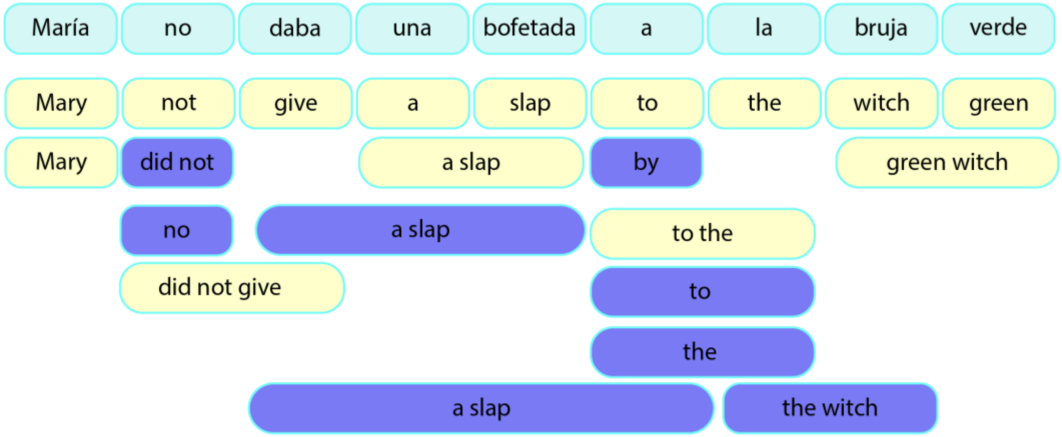
$$\Sigma_{m=1}^{M} \lambda_m \, . \;\; h_m \, (\mathrm{T}, \mathrm{S})$$

The log-linear model is a formula that aids “to combine several models and compute an overall score for each translation hypothesis” (Perez, 19). According to this formula, *argmaxT* stands for the translation with the highest score; *M* indicates the number of models that the decoder uses. Then, as Perez states, “*λm* refers to the weight of the *m*-th model according to its importance, and *hm* (T,S) is the logarithm of the probability provided by the m-model or its value.”

Regarding the decoding phase (refer back to Figure 1), the decoder works under a search-problem principle. In other words, given an input sentence, the decoder works on collecting only the target language hypothesis with the highest probability scores and dropping the ones with less probabilities. Recent STMs employ a beam-search decoder, which “is just an arbitrary number of hypotheses” (Hearne and Way, p.222). Along the translation process, this kind of decoder stores a finite number of translation hypothesis (10,000 hypotheses, for example). They also work on an expansion basis, meaning that a newly translated phrase with a higher probability would replace the previous stored translation in a new stack, as the new translation has greater probabilities. The reason for this is also to give more opportunities for other hypotheses to compete and to discard the much poorer translations.

Another important characteristic of the decoding process is that these phrase-based engines have a higher tendency to choose longer phrases rather than words. This explains why in Figure 3 the yellow blocks represent those phrases preferred by the decoder. Rather than choosing “did not” and “give,” (purple blocks), it chooses the entire phrase “did not give” (yellow block). Cattelan (Phrase-based) affirms in one of his seminars that this is due to the fact that the larger the blocks, more meaning they convey, and then more useful they are for a PB-SMT.

Figure 3. The Decoding Phase in a PB-SMT. Example taken from Koehn, 2004.

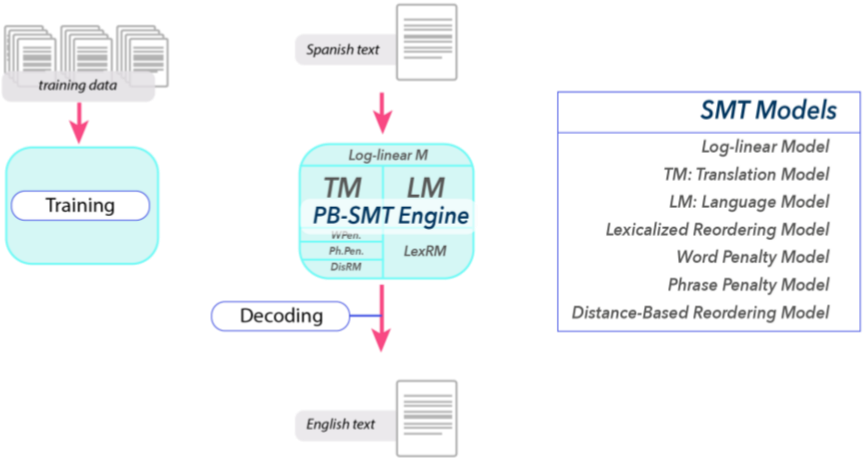


As seen in Figure 3, “not,” “did not,” and “no” are dismissed in the first three stacks and “did not give” remained as the output sentence. The same happens with the input “bruja” “verde”. The decoder first translates it as single words “witch” “green.” However, assuming that the log-linear algorithm applied to this engine uses a lexicalized reordering model to validate the source and target word alignments, the decoder searches for the most likely structure in the target language. In this case, the reordering and language models work together and help to condition the resulting hypothesis. In fact, adjectives preceding nouns are more likely in English grammar

3.2.1 PB-SMT Components

As stated in the Introduction, this research aims to ease the understanding of the theory that underpins MT technologies, so that more translators engage and benefit from its use. Thus, as supported by Kenny and Doherty, we will focus only on “What human translators need to know,” covering only the fundamental components of a PB-SMT. Despite the fact that many ‘features’ or models can be employed in a PB-SMT (see Figure 4), we will only cover the language, translation and log-linear models (the last one was just explained in 3.1).

Figure 4. The Architecture of a PB-SMT System



3.2.2 Language Models P(T)

Language models are trained on a corpus of the target language. In case the translator is dealing with a language pair that has little linguistic resources, the translation side of the parallel corpus can be used, too. Regarding the purpose of the language model *P(T),* Hearne and Way (2011) indicate that it is to compute the probability of *T* to be a correct sentence in the target language. In other words, it works as a filter. It filters out those segments that are not likely to be found in the target language. For this, during the training process, the system uses n-gram models[10]. Perez explains that,

> "these models predict the likelihood that a word follows another word or segment. For example, in the bigram language model, the probability that a word comes after another word is measured. This is computed by dividing the number of occasions in which both words appear one after the other by the total number of appearance of the word in the corpus."

Based on this, Perez explains that n-gram models learn the distribution of words from the corpus, assign them probabilities, and later upon unseen strings, they are scored accordingly, and if the likelihood probability is high, they are kept. To illustrate, in a unigram model, the probability of sentence (1) can be computed like this:

[10] Sketchengine.co.uk defines n-grams as "sequences of words. A unigram is one word, a bigram is a sequence of two words, a trigram is a sequence of three words, etc."

“Marie is the smartest in class, the smartest in school.” **(1)**

Table 1. An example of a sentence using a unigram model

en-word	P(en-word)
marie	0.1
is	0.1
the	0.2
smartest	0.2
in	0.2
class	0.1
school	0.1

Note: en-word refers to “English word” and p(en-word) stands for the probability of that English word.

In a sentence of ten words, like (1), “marie” appears one time in 10 words, “smartest” appears two times in 10 words, and so on. In probabilistic terms, the equivalent of that previous statement is that “marie” occurs one in ten (0.1), whereas “smartest” occurs two in ten (0.2). Then, as supported by Kenny and Doherty, the algorithm of probability can also compute the probability of complete sentences by multiplying all unigram probabilities. See the following examples.

“Marie is the smartest in class.” **(2)**

P(marie)*P(is)*P(the)*P(smartest)*P(in)*P(class)

= (0.1) *(0.1)* (0.2)* (0.2) * (0.2) * (0.1)

=0.000008 (or 8 in 1.000,000)

“Marie is.” **(3)**

P(marie)*P(is)

= (0.1) * (0.1)

= 0.01 (or 1 in 100)

As illustrated in sentence (2) and (3), (3) is much more likely than (2). This points out one of the weaknesses of n-gram models, since they tend to favor shorter sentences. In fact, Microsoft states that since “it only translates words within the context of a few words before and after the translated word. For small sentences, it works pretty well. For longer ones, the translation quality can vary from very good to, in some cases, borderline nonsensical."

Another limitation of n-gram models is the fact that the model assigns the same probability to non-plausible[11] sentences. For example, if we compare sentence (2) with sentence (4), we obtain the same probability: 0.000008 (or 8 in 1.000,000), in spite of the fact that (4) is not an English sentence.

"is in class smartest the Marie." **(4)**
P(is)*P(in)*P(class)*P(smartest)*P(the)*P(Marie)
= (0.1) *(0.2)* (0.1)* (0.2) * (0.2) * (0.1)
=0.000008 (or 8 in 1.000,000)

To cope with the fact that LMs are not linguistically knowledgeable, Kenny and Doherty suggest to use several language models (unigram, bigram, trigram, up to 7-gram models) to "combine the strength of shorter, more flexible n-grams with longer, more context-sensitive n-grams" (p.281).

Finally, another characteristic of these n-gram models is how they manage to compute unknown (O.O.V.) and rare words. Rare words refer to the words with less instances (representation) in the translation model; whereas, unknown words are literally inexistent words in the training data. To explain this, let's use a bigram model. See Table 2.

Table 2. An Example of a Bigram Model

en-word	**P(en-word)**
marie is	1/1
is the	1/1
the smartest	2/2
smartest in	2/2
in class	1/1
class the	1/1
the smartest	1/1
smartest in	2/2
in school	1/1

Note: *en-word* refers to "English word" and p(en-word) stands for the probability of that English word.

[11] Non-plausible means that the sentence is not linguistically correct.

The probability of sentence (5) in a bigram is calculated by “diving its frequency in our corpus by the frequency in our corpus of the first word in the bigram.” (Kenny, p.280). See sentence (5).

P (Marie is the smartest in class) **(5)**

$= P(is|Marie)*P(the|is)*P(smartest|the)*P(in|smartest)*P(class|in)$

$= (1/1) * (1/1) * (2/2) * (2/2) * (1/1)$

$=1$

Sentence (5) has a probability of 1. However, see what happens when adding an unseen string (6) with an unknown word to the language model.

P (Marie is the tallest in class) **(6)**

$P(is|Marie)*P(the|is)*P(tallest|the)*P(in|tallest)*P(class|in)$

$= (1/1) * (1/1)* (0/1)* (0/1) * (1/1)$

$=0$

In bigram models, we can compute the probability of a sentence by the probability of each bigram. In this case, since the bigram (in|tallest) is not found in the bigram model or in the training data, then the probability of that sentence is zero. Hearne and Way argue about a way to differentiate between rare and unknown bigrams. They ponder on adding more training corpus, but this solution will “yield to limited returns as longer n-grams are used.” Thus, new techniques are employed such as giving greater weight (a level of importance) to rare n-grams. This helps to distinguish unknown n-grams (which score zero likelihood) from rare n-grams (which score a low probability score, but different from zero).

3.2.3 Translation Models P(S|T)

Translation models also work under a statistical principle. Its probability formula P(S|T) stands for the probability of a target segment to be the equivalent of the source segment

(Perez, p.10). To calculate this, just like language models, translation models use n-grams, together with word and phrase alignment models.[12] Koehn (Europarl) explains that,

> “When translating a sentence, source language phrases (any sequences of words) are mapped into phrases in the target language, as specified by a probabilistic phrase translation table. Phrases may be reordered, and a language model in the target language supports fluent output.”

In other words, since the bitexts are already aligned at a sentence level, the n-gram models break the parallel corpus, first, into words and then into phrases of varying length (as seen in 0, researchers suggest using up to 7-gram models). After that statistical analysis, each of these aligned biwords and biphrases get scored and stored in a so called Phrase Table. To illustrate, see sentence (7), together with Table 3.

Table 3. Example of a Phrase Table, Following a Trigram Model

It is raining cats and dogs (7)

en-phrases	es-phrases
it is raining	llueve
~~is raining cats~~	está lloviendo gatos
~~raining cats and~~	está lloviendo gatos y
cats and dogs	a cántaros

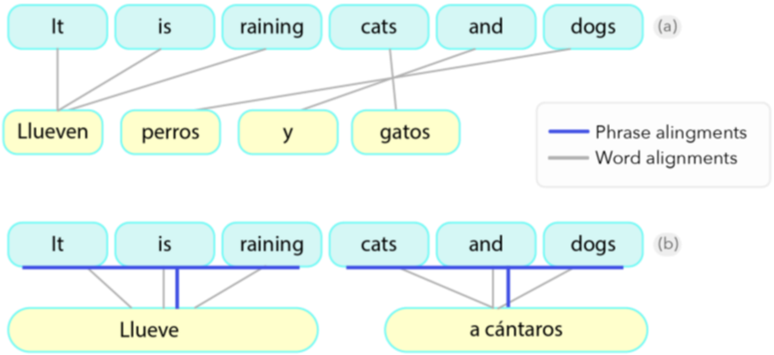
As you can see in Table 3, a phrase table is the representation of the source language and its corresponding equivalent in the target language. People often call the phrase table a dictionary[13]; however, this is a misconception since not all alignments are consistent. Figure 5, is an idiomatic sentence that challenges phrase translations, in semantic terms. This example originates from the German version given by Philipp Koehn[14] (2015).

[12] Word Alignment models were introduced in Brown et. al 1990:80-81. The state-of-the-art PB-MT engines employ phrase alignment models. For more information, refer to Koehn, Och et all 2003.

[13] Germann, U. refers to phrase tables as dictionaries. “Phrase-based statistical MT translates by concatenating phrase-level translations that are looked up in a dictionary called the phrase table.

[14] In this paper, Koehn uses the German sentence “Es schüttet aus Eimern,” which translates to “It pours from buckets.”

Figure 5. Representation of Word and Phrase Alignments (A unigram and a trigram model)



In sentence (a), the alignments of "it," and "is," are not completely consistent. In the scenario of entering an unseen string, the engine will misleadingly retrieve probabilities assigned to the link of "It" with "llueven." In fact, Galbrun classifies these kinds of issues by the name of distortion and fertility, where "*distortion, i.e.,* the fact that the translations of some words may be swapped*,* and fertility, the fact that one word is not always translated into exactly one word in the other language". To fix this, Koehn suggests to use a number of n-grams, making it possible to associate "It is raining" with the one-word action in Spanish "llueve." For more advanced users, Galbrun (2009) suggests to reweight the models and retrain the engine to fix this issue (the tuning phase). I.e., Galbrun urges translators to add a tune-up data set to reach more optimal results. This can be achieved by crawling techniques in order to find in-domain resources in line with the purpose of a custom engine. It is important to highlight that by adding more related-material into an engine, the frequency of those valid alignments will increase. Platforms such as Microsoft Translator Hub, indeed, suggests to use an auto tune configuration, but in case of wanting to have more control over the quality of the engine, they recommend adding a custom tune-set throughout the training of several engines[15].

[15] In its manual, Microsoft Translator Hub states, "Should you decide to create your own Tuning files, make sure they are a random set of sentences across domains if you wish to create a general purpose translation model."

3.3 Neural Machine Translation

PB-SMT and NMT systems share many characteristics; they are data-based and their generic core structure lacks of linguistic knowledge. In comparison to PB-SMT systems, they also face similar challenges, but a few of them have been successfully overcome by NMT engines. Regarding their differences, the most significant one is that they (NMT engines) can translate complete sentences, instead of phrases or chunks. In addition, they differ in terms of their translation process and components. First, NMT systems consist of three recurrent neural networks, which activate in different stages of the translation process (See Figure 6).

Figure 6. The Translation Phases and Corresponding States of a NMT

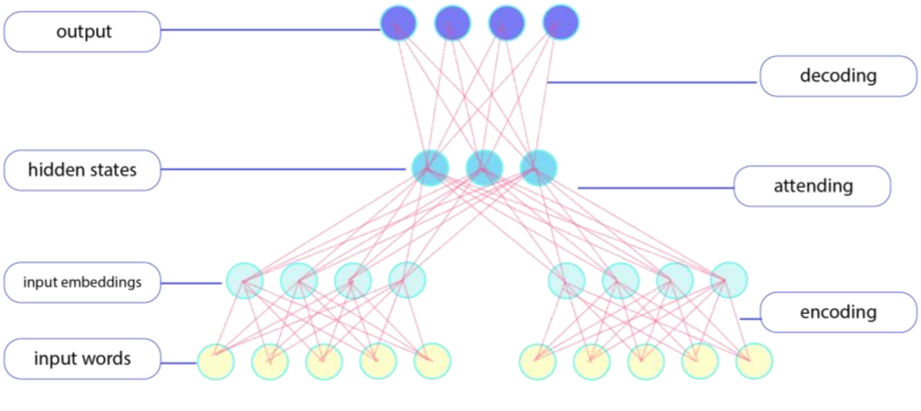
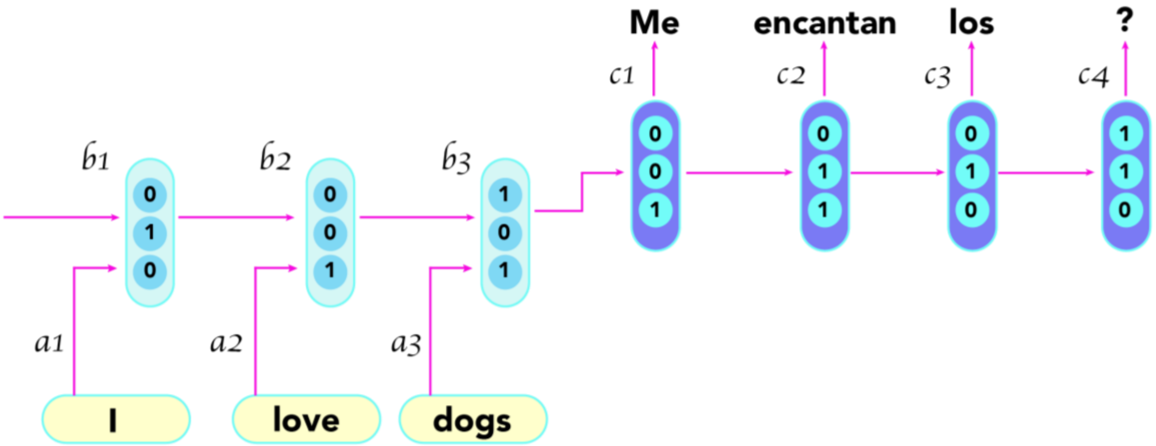


Figure 6 clearly illustrates the translation process, which consists of three main phases: encoding, attending and decoding. The encoding network encodes words into vectors. This encoding phase, in which words shift to vectors (a1, a2, a3 Č b1, b2, b3), is also called word embedding. According to Microsoft (2018), once they are converted into vectors, they are placed into "a 1000-dimension vector", inside multi-layer neurons. In other words, neurons have multiple layers that store the different vector representations of meaning along the encoding phase; later on, upon the completion of the encoding phase, an attention network weights out an average weight representation and points out to the decoder which of those representations or hidden layers should be paid more attention to show how words within a sentence are encoded, and how later on, provided that the hidden state of b3 is scored the highest by the attention network, the decoder would then use the output text of b3 to produce the next target word.

Figure 7. Word Embeddings use Float Numbers[16]

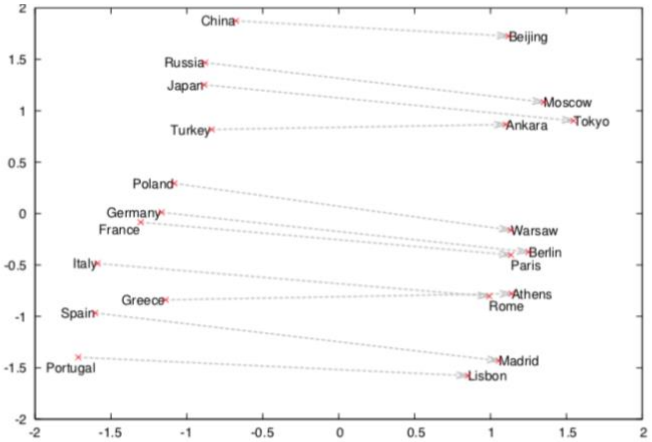


In addition, the encoding phase or transformation of source sentences into sequences of vectors is done through the use of float numbers[17] or else a weight per each of these vector words. The main objective of this process is to get a representation of the word, and based on that, map out the meaning of source words into a vector space, stored by the networks themselves. Figure 8 depicts how word embeddings are mapped out within the vector space. They are distributed on a vocabulary similarity basis. In this specific example, the Recurrent Neural Networks have automatically grouped these words in two groups: countries and cities.

[16] The float numbers entered here are used only for illustration purposes. The original example can be found in Britz, 2016

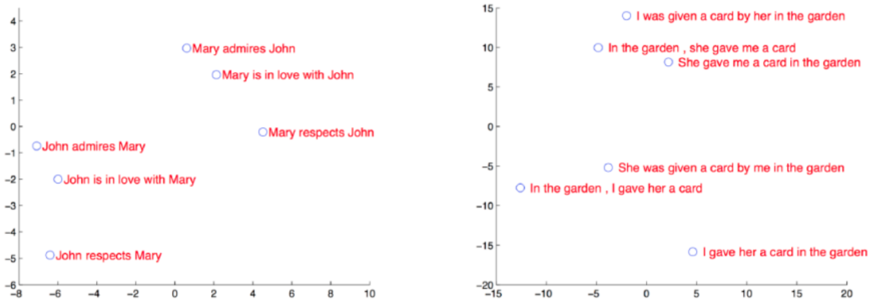
[17] A number in scientific notation with no leading 0s is called a Normalised Number: 1.0 × 10-8. A non-normalised form: 0.1 × 10-7 or 10.0 × 10-9. It can also represent binary numbers in scientific notation: 1.0 × 2-3. Computer arithmetic that supports such numbers is called Floating Point.

Figure 8. Distribution of Words Within the Vector Space. Example taken from Mikolov,T. et al. 2013.



Another important characteristic of NMT systems is that they translate full sentences, rather than 3 to 7 n-gram phrases like PB-SMTs. Thus, neural networks have the capacity to store meaning at a sentence level. See Figure 9. In the first graphic, source sentences are sorted by the agent who is performing an action. In this case, all sequences starting with Mary are grouped together; whereas all sentences performed by John are set farther in the vector space.

Figure 9. The last encoded state of neurons or the last hidden-state of source sentences (image taken from Sutskever et al., 2014, as cited by Haddow)



After the encoding phase, the decoder aims at finding the full translation of a sentence with the highest likelihood. As mentioned before, to accomplish this, the attending network performs a weighted sum value, “a weighted combination of all the input states” (Britz, 2016). According to Marcello et al. (2016), “the attention model informs the decoder about the encoding hidden states corresponding to the next target word.” In other words, two different tasks are carried out almost simultaneously to generate a translation. On one hand, the attention network uses the final representation of meaning or “the final source hidden states” (Haddow, p.29) and the last translated word to determine the word order of the translation. On the other hand, the decoder uses a beam search algorithm to find the best representation of the entered data.

Today, the most powerful feature of the state-of-the-art NMT is its learning skills. Learning upon corrections occurs thanks to the fact that the engine is equipped with full sentence storage capacity and the use of recurrent neural networks (RNN). As Cattelan (Phrase-Based) explains, they learn by means of “input and corrections.” In the scenario of a translation editor, when the translator corrects an inaccurate translation and confirms the correct segment, because this engine is cyclic-driven, this new data would run back to the vector space and sort the float numbers in accordance to the correct input provided by the translator. This does not only condition inaccurate translations to not replicate in the future, but it also makes possible instant retraining without having to add new data sets. To support this, Microsoft (2018) states that “Once all words have been encoded one time into these 1000-dimension vectors, the process is repeated several times, each layer allowing better fine-tuning of this 1000-dimension representation of the word within the context of the full sentence.” This infinite capacity to refinement of neurons promises great expectations as it allows infinite opportunities of improvements over time.

Moreover, it is important to highlight the names and functions of the main algorithms that play a key role in the translation process. For computing statistical hypothesis, NMT systems use a softmax function, which basically works as a converter of vector weights to probabilities. According to Haddow, this conversion reduces “cross-entropy” (p.19), making the statistical data easier to interpret for the decoder. Another key algorithm is the one used by the beam searcher. This beam search takes place during the decoding phase. The algorithm focuses on “coverage penalty and length normalization.” According to Wu et al. (2016), ”without some form of length-normalization regular beam search will favor shorter results over longer ones on average since a negative log-probability is added at each step, yielding lower (more negative) scores for longer sentences.” These enhancements to NMT generic core (encoder-decoder) structure has been enforced to overcome the decrease of quality when translating long sentences. In the case of coverage penalty, it aims at solving over-translation and under- translation issues. In *Modeling Coverage for Neural Machine Translation* (2016)," researchers refer to coverage penalty as a mechanism to not overlook linguistic data.

In the form of a summary, the table below synthetizes all the information we have covered so far in the Literature of the Review regarding the basic components and functions of the core elements of PB-SMT and NMT technologies. The table also helps to distinguish the main similarities and differences between them. Learning this distinction would be essential to understand the following chapters of this research.

Table 4. Comparison chart between PB-SMT and NMT

	PB-SMT	NMT
Type of technology	Data-driven No linguistic knowledge required.	Data-driven No linguistic knowledge required.
Translation process	Training, tuning and decoding	Encoding, attending and decoding
Major components	Translation, language, n-gram models	Word embedding, recurrent neural networks
Translation quality	Takes in phrases of varied length input. Translation may look too mechanical.	Takes in sentences. According to Microsoft, NMT produces "more fluid and human-translated looking translations."
Incremental learning, through post-editing	Offline training	Online training
Algorithm of probability implemented to get the translations with the highest likelihood.	Log-linear	Beam search
Capacity to adapt to domain	Yes, through the use of in-domain corpora, specific terminology and retraining.	Yes, through a context analyzer, adaptive phrase table and language model.[18]

[18] According to Marcello et al. (2016), these specific core elements have been implemented in MMT systems.

Translates unknown and rare words	Yes, issues in the output text are easier to identify and post-edit. The PB-SMT does not translate them.	Yes, but it could be counterproductive as this system is fluency-driven, making post-editing more time consuming. The response of NMT is unexpected.
Performs well in in-domain scenarios	Yes, especially with words that have multiple definitions.	Yes, especially with words that have multiple definitions.
Performs well in out-of-domain scenarios	No	No[19]
Highly dependent of the quality of the training data	Yes. Koehn supports that “doubling the amount of training data gives a fixed increase in BLEU scores.[20]	Yes, and it is more sensitive to grammar inaccuracies.
Struggles with long sentences	Yes, especially because longer sentences have less probabilities than short sentences to be found in the SMT Models	Yes, according to Koehn, NMT engines “do comparably better up to a sentence length of about 60 words.”
Performs well in all language combinations	Yes, in many more pairs since more research has been dedicated to this technology.	Not yet. Koehn affirms that the most reliable pair combination is French>Spanish.
Available platforms or toolkits to build an engine	KantanMT, LetsMT, MTradumática, Microsoft Translator Hub, Moses.	MMT, Nematus, TenserFlow
Current commercial end-user applications	Bing Translator, Google Translate, etc.	Lilt, MateCat, Google Translate, DeepL, etc.

[19] Philipp K. (2017) states that “NMT systems have lower quality out of domain, to the point that they completely sacrifice adequacy for the sake of fluency.”

[20] As cited by Koehn et al. (2017).

4. Methodology

After having learned the fundamentals of data-driven engines, this section attempts to propose an approach to building a custom PB-SMT and an NMT in low-resource linguistic settings such as ES_CR Č EN. In order to avoid any confusion, the methodology is structured in two sections. 4.1 deals solely with the process of training a PB-SMT engine. In section, 4.2, we will attempt to train a NMT engine using the same datasets already prepared in 4.1. Bear in mind that along the process, the only steps that vary are related to the training steps because of the different platforms used and the refinement procedures due to their unique technological features. This explains why a few steps related to data preparation are omitted or slightly overlap along the methodology.

4.1 Training a PB-SMT engine in KantanMT

To build a custom engine, it is necessary to define its translation scope since researchers have demonstrated in many studies that in-domain engines outperform generic ones. In fact, they argue that it is best to train a suite of incremental engines that would have the capacity to adapt to different domains.[21]

Under that premise and the Objectives of this research paper, we checked several institutional websites in search for commonalities within the public sector of Costa Rica. In fact, most of their web content (laws, regulations, frequently asked questions, forms, manuals and digital procedures) share many linguistic characteristics and are intended for the same kind of audience. Another aspect taken into consideration was a study from 2016, in which the Ministerio de Economía, Industria y Competitividad (MEIC) reports a total of 69 Costa Rican institutions and 2044 administrative procedures available online. See the complete statistics report on the online procedures available per public institution in Appendix 2. Due to time constraints and the aforementioned aspects, it was decided to

[21] As cited in Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora, Lumeras and Way (2017) state that "It is well-known that MT systems work best when tested on data that is very similar to the corpora on which they are trained.”

build an engine that would be able to translate most of the content found in the Registro Nacional[22] (National Registry) and MEIC[23]. Both sites store texts such as intellectual and industrial property general information, registration of societies, companies, inventions, patents, contracts, etc. In the ambition of appealing foreign investment, having this particular website available in English would be a great asset and starting point to approach the complete translation of the rest of content that comprises the websites of Registro Nacional and MEIC.

4.1.1 Collecting and preparing the training data

Training data, as defined by TAUS refers to "the set of sentences selected during the process of setting up a statistical machine translation workflow used to train or customize an engine for a specific domain or language pair." The open source OPUS Corpus is one of the most recognized repositories of aligned translations. Although it stores a large number of corpora, the most relevant ones to the domain at stake were the News Commentary 11, EU Central Bank and the Directorate-General for Translation. In addition to the factor of relevance to domain, these corpora were selected over the rest of corpora available because of the use of formal tone and normative Spanish and English.

Regarding the data collection task[24], it simply entailed downloading the TMX files ("Upper-right triangle: download translation memory files", in the Opus) and the raw, non-tokenized, monolingual files in English. By using Olifant[25] a translation memory editor, it was possible to reverse the TMs and to remove possible errors. To do that, it was necessary to first create a new TM repository with English as the target language and Spanish as the source one. Olifant automatically reverts the language pairs when importing the translation entries in a new TM repository. Through the function "Flag entries," we detected a few invalid inconsistent translations and removed them. See an example of this process in

[22] http://www.registronacional.go.cr

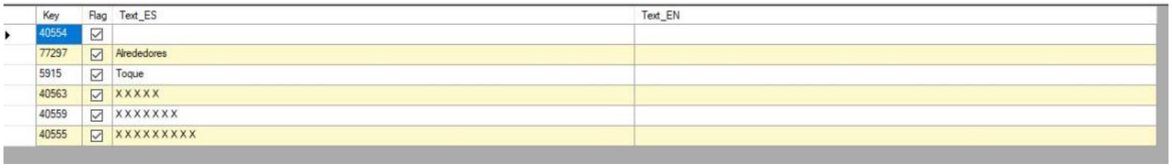
[23] https://www.meic.go.cr/meic/

[24] "MT for Everyone" is a video tuturial series that includes all processes followed in this methodology.

[25] http://okapi.sourceforge.net/Release/Olifant/Help/index.html

Appendix 3. After this, the translation memories of the News Commentary 11 and EU Central Bank were ready to be added to the engine.

Figure 10. Invalid empty target segments in the original TMs, from the Opus Corpus



Key	Flag	Text_ES	Text_EN
40554	☑		
77297	☑	Alrededores	
5915	☑	Toque	
40563	☑	XXXXX	
40559	☑	XXXXXXX	
40555	☑	XXXXXXXXX	

In the case of the DGT corpus, KantanMT strictly accepts files of a maximum of 512 MB. Thus, splitting the translation memory was handled manually with Notepad++[26]. To open the file, it was necessary to use a CPU of 16 GB of RAM; otherwise, computers featuring 8 GB or less would show an error message, making it impossible to proceed with this task. To ease the management of this large translation memory, the original TM (784,5 MB) was split in translation memories of up to 3,010,101 lines. However, depending on the length of the strings, the size of the files varies. Table 5 summarizes the resulting six files of the DGT TM after the splitting operation.

Table 5. Resulting Files from the DGT TM, and their Corresponding Sizes

DGT_es_en_01.tmx	128,3 MB
DGT_es_en_02.tmx	154,9 MB
DGT_es_en_03.tmx	146,1 MB
DGT_es_en_04.tmx	151,6 MB
DGT_es_en_05.tmx	139,1 MB
DGT_es_en_06.tmx	33 MB

Additionally, to avoid any human mistakes, every single file was run through TMX Validator[27], an open source tool to check that all files followed the Translation Memory eXchange format[28]. After this task, all DGT translation memories were ready to train the translation model. See, in Appendix 4, the TMX Validator in action.

Regarding the training data for the purpose of training the language model, the target side of the translation memories was used. The data preparation of the mono texts entailed

[26] https://notepad-plus-plus.org/download/v7.5.6.html
[27] https://www.maxprograms.com/products/tmxvalidator.html
[28] http://www.ttt.org/oscarstandards/tmx/tmx13.htm

downloading the raw files from the Opus Corpus in English, and merging all corpora in one single plain text file, in UTF-8, under the name of "source.utf8.trg.mono"

4.1.2 Training the PB-SMT engine

The customization of a statistical engine was developed in the SAAS platform named KantanMT. After running a few experiments in other commercial platforms such as Microsoft Translator Hub (which supports languages in low-linguistic settings), KantanMT proved to be a superior tool. Among a few of the decisive factors include: training speed, files size capacity, integrated tools for assessment and advanced refinement features. Based on a series of experiments carried out in both platforms with some of the already mentioned training datasets (in 4.1), the following are some interesting findings:

¶ KantanMT takes approximately 30 minutes to complete the training of 356,934 MB, whereas Microsoft Translator Hub takes around 2-3 days to complete the same kind of training.

¶ Another aspect is file management capacity. While MTHub limits the import of individual files up to 100 MB. KantanMT allows up to 512 MB per file.

¶ Regarding the assessment and analysis features, KantanMT provides integrated tools and statistics on F-Measure, BLEU, TER scores and others to evaluate the quality of the training data. MTHub, on the other hand only offers the BLEU score.

Having decided on the appropriate platform to carry out the experiment, it was necessary to first create a client profile (It is possible to gain free access through a student subscription as well). After creating a new engine (on the Dashboard, click on "New"), a wizard window guides you throughout the set-up process. See Table 6 for more details on the entered settings.

Table 6. Training Settings of System 1

Name: System 1	Under a student subscription, the number of engines is limited up to 5, and only 1 job can be run at a time.
Engine type: Statistical MT	Neural MT is unavailable under a student subscription.
Source Language: Spanish {es}	To avoid any inconsistencies in the translation memories, we decided to use the ISO language standards {es}, instead of {es-cr}, throughout the entire project.

Target Language: English {en}	
Library: General (English ė Spanish 2) Source: 6.851.405 Target: 7.318.088	Kantan offers a wide range of domains to choose from, including financial, medical, automotive, legal, etc. The general domain was chosen since the target domain addresses end-users in an informative manner. By choosing a general Kantan library, we aimed to cover those language gaps that the current training datasets might feature regarding the use of common daily language.

After having chosen the engine settings, we clicked on the Training tab of the dashboard to add our training data. Dragged and dropped the exceptionally well cleaned training data onto the training board. (See Appendix 5. Checklist of training data before building a PB-SMT in the KantanMT platform). Table 7 shows the list of corpora used to train the statistical models. Both, parallel and monolingual corpora are dropped onto this window. The system automatically recognizes the language codes and combinations.

Table 7. Training Dataset I

Name	Size
DGT_es_en_01.tmx	122 MB
DGT_es_en_02.tmx	148 MB
DGT_es_en_03.tmx	139 MB
DGT_es_en_04.tmx	145 MB
DGT_es_en_05.tmx	133 MB
DGT_es_en_06.tmx	31 MB
ECB_es_en.tmx	46 MB
News_es_en.tmx	91 MB
source.utf8.trg.mono	409 MB
General 2 (Source WC: 6,851,405, Target WC: 7,318,088)	

After having successfully uploaded all documents, the next step was to click on “Build” to start the training phase. Note that there are other options such as “Adapt” and “Rule Editor” that can be applied after the first training to enhance the translation quality. In the first case, the adaptation feature helps you add corrections to your engine without having to retrain or create a new one. Simply add your post-edited translations into the Translation tab to make corrections. In the case of the Rule Editor, KantanMT offers hybrid functions by combining rule-based technologies with statistical ones. In other words, the user can apply “its own pre-processing and post-processing rules.” Since our scope is to use only data-driven approaches, we disregarded the last function in this methodology. However, for further reference and assistance, the user can always access the Training Help[29].

The training process took 12:32:43. Once the training is completed, KantanMT sends an email notification to confirm its status, together with a report that indicates the automated metrics scores.

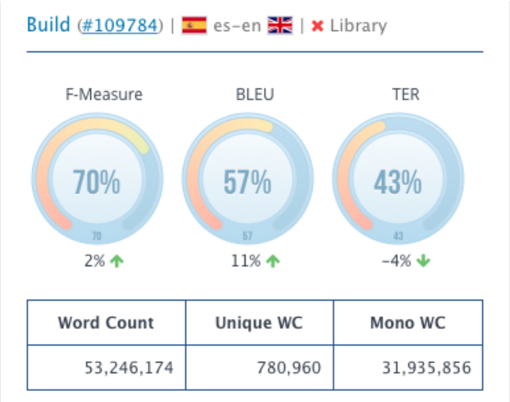
4.1.3 Measuring System 1

KantanMT is well-equipped to support continuous improvements to custom systems. In fact, their "evolutionary concept" is consistent with its integrated analysis tools named BuildAnalytics and KantanAnalytics. These tools offer a wide range of insights about the engine right after the training is completed and should help to get a first appraisal of the engine's performance. At first sight, BuildAnalytics indicates that the "engine shows good understanding of your target domain and language." Now, for further analysis, the performance of System 1 will be appraised in accordance with additional reports such as the automated metric scores, the rejected segments from the training data, the gap analysis and the output text of unseen strings. A careful examination of the current engine's performance will lead to the design of an objective action plan to improve System 1's metrics, and hopefully quality as well.

As seen in Figure 11, KantanMT provides different word counts and three main scoring metrics: BLEU, F-Measure and TER.

[29] https://app.kantanmt.com

Figure 11. Automated Metrics of System 1

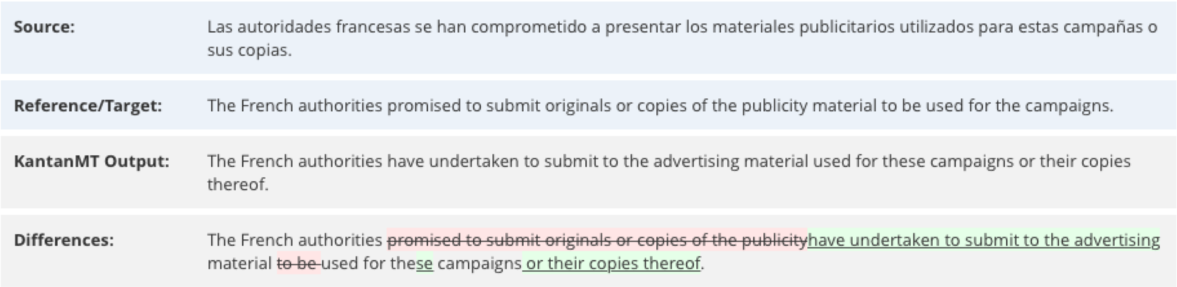


Build (#109784) | es-en | ✖ Library

Word Count	Unique WC	Mono WC
53,246,174	780,960	31,935,856

The BLEU score stands for Bilingual Evaluation Understudy. This metric was designed to replace the expensive cost and subjectivity of developing human evaluation assessments. Kantan Help guide explains that "the BLEU metric measures how many words overlap in a given translation when compared to a reference translation, giving higher scores to sequential words." To keep it short, it measures MT's output fluency.[30] Based on KantanMT's BLEU score standards, an engine should aim for high, up to 100% scores. As seen in Figure 11, the training on 53.246.174 million source words resulted in 57% of BLEU score. From the TAUS's 4-point scale of fluency, the output text of System 1 is deemed as disfluent. See Figure 12, an example of the dystrophy of System 1 to translate.

Figure 12. Segment 29 from the test set provided by BuildAnalytics. This segment scored 5% of BLEU: Incomprehensible Segment.

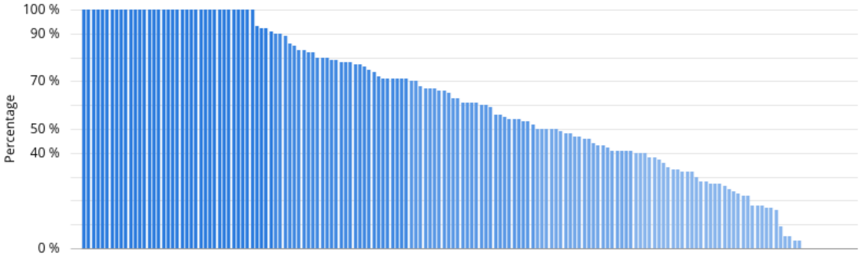


Source:	Las autoridades francesas se han comprometido a presentar los materiales publicitarios utilizados para estas campañas o sus copias.
Reference/Target:	The French authorities promised to submit originals or copies of the publicity material to be used for the campaigns.
KantanMT Output:	The French authorities have undertaken to submit to the advertising material used for these campaigns or their copies thereof.
Differences:	The French authorities ~~promised to submit originals or copies of the publicity~~have undertaken to submit to the advertising material ~~to be~~ used for these campaigns or their copies thereof.

[30] According to TAUS, "fluency denotes to what extent the translation is one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker. Fluency can be evaluated segment-by-segment on a 1-4 scale (Flawless, Good, Disfluent or Incomprehensible)."

As seen in this example, System 1 produces serious mistranslations. On a meaning level, while in the source string it is not reported to who shall the papers be submitted to, in the string produced by System 1, the papers are to be sent to the “advertising material.” The engine also fails by adding new content (the word “thereof” at the end of the string). This addition does not affect much the overall meaning of the segment, and these MT errors are easy to detect in post-editing. Although cases like this one scatter throughout the analysis, there are as well high-quality translations scores. To fully comprehend the BLEU score, Kantan offers a more visual understanding of the overall BLEU score, named Score distribution.

Figure 13. BLEU Score Distribution of System 1



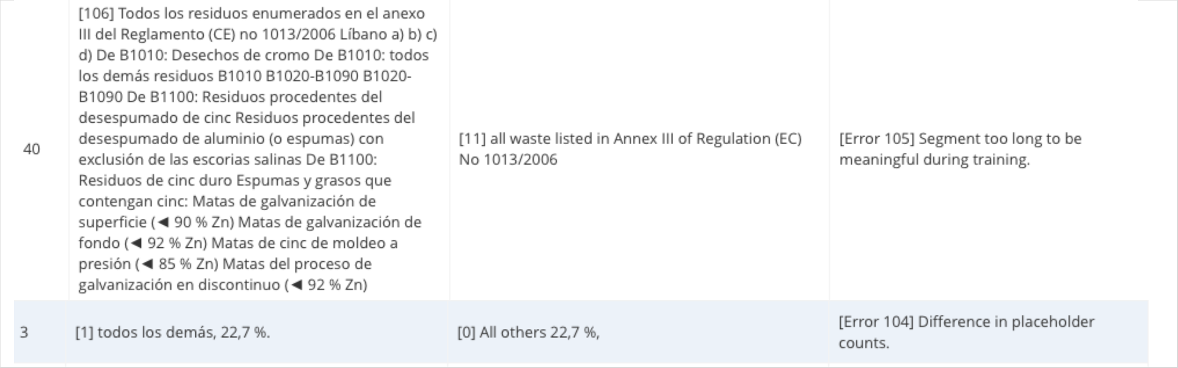
Based on this figure, it is clear that over a third portion of the training data scores more than 40% of BLEU, which is a positive and acceptable score. However, KantanMT suggests improving the engine’s BLEU score up to 60%: Good Fluency. This can be done through the addition of parallel data.

Regarding the TER score, it stands for Translation Error Rate. This metric was designed to predict the post-editing cost of machine translation content to reach publishable quality. In general, KantanMT reports the need of minimal human editing, which is a favorable indicator since all content from the e-government is addressed to end-users, and reducing the post-editing cost would be critical in a large-scale project like this one. Based on the score given by the report on Figure 11, System 1’s TER score should decrease at least 3% to reach “a good quality” qualification (any TER score lower than 40% is better).

Finally, the F-Measure refers to the capacity of an engine to translate content. The Kantan Help guide explains that it “measures how precise KantanMT operates when retrieving words and how many it can retrieve or recall during translation.” Even though the present System 1 features the average F-Measure score (70%), as seen in Figure 11, Kantan reports that System 1 “has below average knowledge of your target domain and language.” This statement is very contradicting to what was previously indicated in the summarized version of the F-Measure Score: “Your engine has an average knowledge of your target domain and language.” Yet this might be due to the number of unknown words[31] detected by the Gap Analysis. However, many of these unknown words are false positives such as “idea,” “India” and “fundamental”[32]. For this, KantanMT suggests to download the file and add them as a ignorewords.txt file in the Training data tab.

Besides the automated metrics, another important indicator to reflect on the quality of the engine, particularly the quality of the training data, is to check the Training Candidates Rejects Report, which informs what training data was actually used. Since KantanMT follows a GIGO principle (Garbage In and Garbage Out), it employs its own cleansers to guarantee that only clean data is used in their systems. In average, a 9% of the training data was rejected. This was due mainly to Error 104 (difference in place holder counts) and Error 105 (segment too long to be meaningful during training). See Figure 14 for examples and Appendix 7 for the complete report.

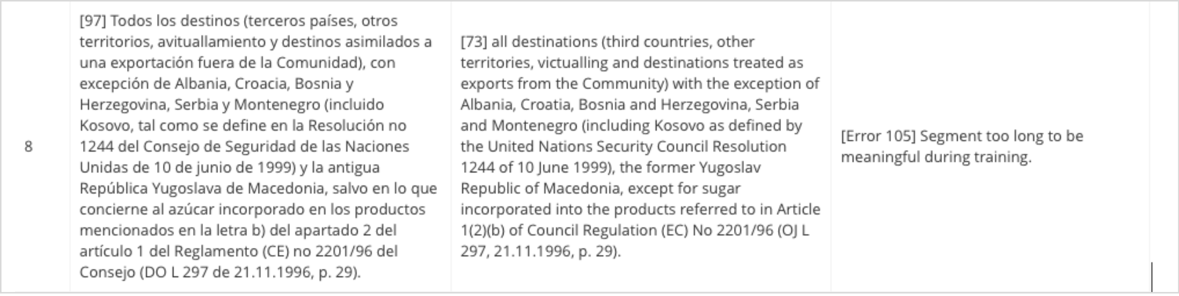
Figure 14. The most recurrent rejected segments of System 1 were Error 104 and 105.



40	[106] Todos los residuos enumerados en el anexo III del Reglamento (CE) no 1013/2006 Líbano a) b) c) d) De B1010: Desechos de cromo De B1010: todos los demás residuos B1010 B1020-B1090 B1020-B1090 De B1100: Residuos procedentes del desespumado de cinc Residuos procedentes del desespumado de aluminio (o espumas) con exclusión de las escorias salinas De B1100: Residuos de cinc duro Espumas y grasos que contengan cinc: Matas de galvanización de superficie (◄ 90 % Zn) Matas de galvanización de fondo (◄ 92 % Zn) Matas de cinc de moldeo a presión (◄ 85 % Zn) Matas del proceso de galvanización en discontinuo (◄ 92 % Zn)	[11] all waste listed in Annex III of Regulation (EC) No 1013/2006	[Error 105] Segment too long to be meaningful during training.
3	[1] todos los demás, 22,7 %.	[0] All others 22,7 %,	[Error 104] Difference in placeholder counts.

[31] Unknown words refer to words that are not found within the training data.

[32] See the complete list of unknown words detected by the Gap Analysis in https://drive.google.com/file/d/1vTiA13b6CHL0EwVozzoH1N4Szg05hMBW/view?usp=sharing



8	[97] Todos los destinos (terceros países, otros territorios, avituallamiento y destinos asimilados a una exportación fuera de la Comunidad), con excepción de Albania, Croacia, Bosnia y Herzegovina, Serbia y Montenegro (incluido Kosovo, tal como se define en la Resolución no 1244 del Consejo de Seguridad de las Naciones Unidas de 10 de junio de 1999) y la antigua República Yugoslava de Macedonia, salvo en lo que concierne al azúcar incorporado en los productos mencionados en la letra b) del apartado 2 del artículo 1 del Reglamento (CE) no 2201/96 del Consejo (DO L 297 de 21.11.1996, p. 29).	[73] all destinations (third countries, other territories, victualling and destinations treated as exports from the Community) with the exception of Albania, Croatia, Bosnia and Herzegovina, Serbia and Montenegro (including Kosovo as defined by the United Nations Security Council Resolution 1244 of 10 June 1999), the former Yugoslav Republic of Macedonia, except for sugar incorporated into the products referred to in Article 1(2)(b) of Council Regulation (EC) No 2201/96 (OJ L 297, 21.11.1996, p. 29).	[Error 105] Segment too long to be meaningful during training.	

In this case, although a clean-up process took place as explained in 4.1.1, it is not possible to develop further cleaning since that would involve actual translation. In a real-life scenario, the translators interested in building custom engines for their clients are expected to provide their own translation files, guaranteeing high quality material. The fact that the training data is not of the highest quality was considered to be a minor risk prior to training the engine, and now that it has become evident that the training data is not completely consistent, it should remain as a minor weakness. The improvement of the training data can only take place over time, by adding more post-edited translations and retraining the engine with training data of higher quality.

A final step of the measuring phase is to run a translation analysis of unseen strings. For this, a source testing text was created (see Appendix 8), which includes four excerpts, arbitrarily selected from the National Registry of Costa Rica and MEIC. Below are the headers of each of the selected texts:

¶ Texto 1: Las preguntas más frecuentes

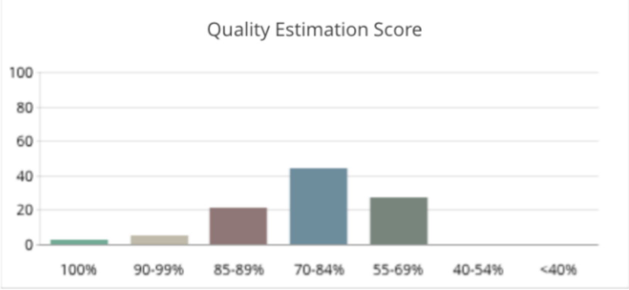
¶ Texto 2: Procedimiento simplificado para extranjeros

¶ Texto 3: Registro Nacional de sociedades

¶ Texto 4: Servicios

In the Translation tab, we added the source text; then, clicked on Analyze. KantanAnalytics gives an estimate of the translation cost. To do this, it works under a translation memory principle, providing the user a segment by segment quality score. The analysis is given in the form of a Fuzzy Match format; thus, it shows the recall capacity of the engine through a matching segment scheme. Based on Figure 15, most words (312 out of 709 words) scored a quality of 70-84%, while the second highest group of matched words (193) scored a quality of 55-69%.

Figure 15. Quality Estimation Score of System 1



In addition to the translation analysis, when executing the corresponding translation job with the Translation option, the engine provides not only the translated output text, but also a file of unknown words. See Appendix 10 for the full report. We found out that 6 out of 20 unknown words are related to specific names of national public institutions such as CrearEmpresa, SETENA, SENASA, INS, Caja Costarricense del Seguro Social and Dirección General de Tributación. To fix these unique translation terms, we are required to add a run-time glossary, which applies search and replace features. The rest of the unknown terms, like verbs (“digitando”, “digitarse”, “indicadas”, “inscriben”) should be added to a general glossary since they have varying forms depending on the grammar tense.

Having fully discussed the insights of System 1 provided by KantanMT, in terms of automated metrics, recall capacity of the engine upon unseen strings, unknown words reports and translations cost, there is still plenty of room for improvement. Based on this appraisal, we will follow an incremental improvement approach, modifying, adding more training data and retraining System 1. Table 8 summarizes the action plan to refine System 1.

Table 8. The Refinement Action Plan of System 1

	Current – Target Score	Means of improvement
BLEU	58% => +60%	More parallel texts
TER	61% => -40%	More parallel texts
F-Measure	71% => NA	Ignorelist.txt
Training data quality	Average 9% of segments were rejected	NA. there is not an immediate solution. It can improve in the long-term.
Unknown words	20	Run-time glossary General glossary
MonoWC	31,935,856 out of 53,246,174	Add distinct training data from the TMX files.

As seen above, the Refinement Action Plan includes the change of the monolingual dataset. In spite of the fact that KantanMT does not provide any insights regarding the monolingual text provided, after contacting the Senior Client Solutions Engineer of KantanMT, Riccardo Superbo, he affirms that KantanMT has not taken iccanto account the current monotext of almost 30 million words, since it interprets them as duplicates from the bitexts. According to this, it will be necessary to add a completely different corpus to train the language model.

4.1.4 Refinement of System 1

Based on the action plan mentioned before, in section 4.1.3, we will describe the specific procedures carried out to gather translations in the combination ES_CR Č EN. In section 4.1.4.2, we applied a crawling mechanism to obtain more in-domain bitexts (the OEMP case), in 4.1.4.3, we applied a different platform to crawl a website in English, which will work to feed the new language model (the WIPO case). Additionally, in section 4.1.4.4, we will describe how the different glossaries were built.

4.1.4.1 Gathering Costa Rican Parallel Corpus

Collecting training data in the combination of ES_CR Č EN was a really challenging task. Unfortunately, Costa Rica neither counts with a strong open data policy, like the EU Open Data[33] nor an open repository of training data, like the OPUS Corpus. Thus, to retrieve public translations of important documents such as the Political Constitution of Costa Rica, the Electoral Code and two resolutions from the Costa Rican Constitutional Chamber, we made use of the master's thesis from the online database of the Universidad Nacional de Costa Rica, SIDUNA[34]. In addition, when possible, the original translators were contacted.

In the case of the Electoral Code, the translator, Gabriela Castro, sent the respective files in .docx format via email. After checking for any misspellings in Microsoft Word[35], we

[33] This initiative was first enforced in 2012. (European Commission, 2011)
[34] http://www.opac.una.ac.cr/F?RN=652173929
[35] Doing this affects positively the quality of the training data, thus the output text produced by the engines.

proceeded to create an alignment project, using the module *LiveDocs* in memoQ[36]. The alignment of the Electoral Code was relatively easy in comparison to the Political Constitution and the Constitutional Chamber Resolutions, since the last ones were retrieved in pdf. format. The pdf files were converted into text, and after, a Regex Clean-up[37] was applied. To convert the files, we used pdftotext[38] (an online tool). Once we obtained these plain texts, we opened them with TextWrangler[39] and executed a few regular expressions to remove corrupted spaces, page numbers, among others. See Table 9 with a list of regular expressions that were applied to most texts.

Table 9. Regex Clean-Up Summary

Find	REPLACE	FUNCTION
^([0-9]{2})$		To delete page numbers of two digits.
página [0-9]* de [0-9]*		To delete page numbers with the following format: "Página 01 de 10."
https?:\/\/(www\.)?[-a-za-z0-9@:%._\+~#=]{2,256}\.[a-z]{2,6}\b([-a-za-z0-9@:%_\+.~#?&//=]*)		To remove any URL.
<[^>]*>		To remove all html tags[40].
\n\d[^\n]*		To remove segments that begin with numbers.
\s+?$		To remove trailing whitespaces.
,\n^ |;\n^	,	To delete incorrect break lines after commas.
		To delete double spaces.
([\n\r]+)		To delete invalid line breaks and return carriages within sentences.
^		To delete spaces at the beginning of paragraphs.

Once the bitexts have been cleaned up, we proceeded with the aligning task. One of the difficulties encountered during this phase was the fact that the translators have added specific annotations to their translations. Thus, the English version featured extra segments that did not match with any source segment. For this memoQ's statistical algorithm to align

[36] https://www.memoq.com/es/

[37] TAUS defines data cleaning, for machine learning purposes, as the process of removing all kinds of formatting, tags and other elements in bitexts.

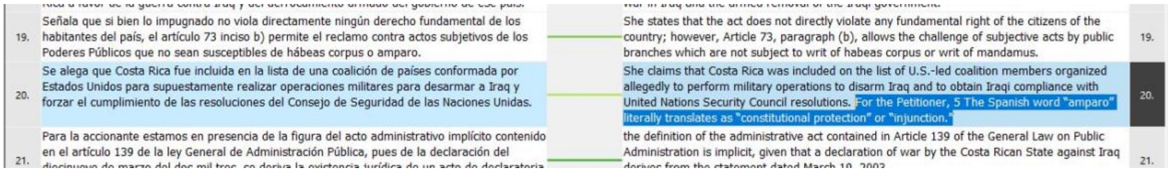
[38] http://pdftotext.com/

[39] https://textwrangler.uptodown.com/mac

[40] A few of these regular expression were taken from the work of Peña, V. Entrenament de motors de traducció automàtica estadísticaentre el castellà i el romanésespecialitzats en farmàcia i medicina.

failed several times and created a great amount of noise. Figure 16 makes evident how the translator explains the target reader what the term “amparo” means in the Costa Rican, legal context. Examples like these scattered over the CR parallel corpora, making the alignment task really time consuming.

Figure 16. Translators have adopted footers to offer more detailed-information about the definition of specific terms.



	Rica a favor de la guerra contra Iraq y del derrocamiento armado del gobierno de ese país.	war in Iraq and the armed removal of the Iraqi government.	
19.	Señala que si bien lo impugnado no viola directamente ningún derecho fundamental de los habitantes del país, el artículo 73 inciso b) permite el reclamo contra actos subjetivos de los Poderes Públicos que no sean susceptibles de hábeas corpus o amparo.	She states that the act does not directly violate any fundamental right of the citizens of the country; however, Article 73, paragraph (b), allows the challenge of subjective acts by public branches which are not subject to writ of habeas corpus or writ of mandamus.	19.
20.	Se alega que Costa Rica fue incluida en la lista de una coalición de países conformada por Estados Unidos para supuestamente realizar operaciones militares para desarmar a Iraq y forzar el cumplimiento de las resoluciones del Consejo de Seguridad de las Naciones Unidas.	She claims that Costa Rica was included on the list of U.S.-led coalition members organized allegedly to perform military operations to disarm Iraq and to obtain Iraqi compliance with United Nations Security Council resolutions. For the Petitioner, 5 The Spanish word "amparo" literally translates as "constitutional protection" or "injunction."	20.
21.	Para la accionante estamos en presencia de la figura del acto administrativo implícito contenido en el artículo 139 de la ley General de Administración Pública, pues de la declaración del diecinueve de marzo del dos mil tres, se deriva la existencia jurídica de un acto de declaratoria	the definition of the administrative act contained in Article 139 of the General Law on Public Administration is implicit, given that a declaration of war by the Costa Rican State against Iraq derives from the statement dated March 19, 2003	21.

4.1.4.2 Oficina Española de Marcas y Patentes (OEMP)[41]

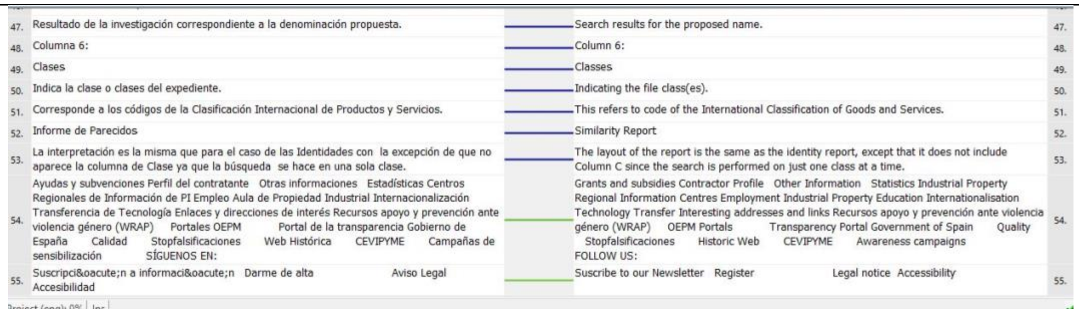
We also crawled the OEMP in search of more in-domain translations. Through optimized engine searches in Google.com (insite: “patent” “patente”), the OEMP tops the list of sites with bilingual content. The first positive indicator that was thought to point out to a potential site was its map site. Additionally, this site pointed to external sources such as the National Registry of Costa Rica. Among other pages that resulted from the previous search was the World Intellectual Property Organization[42] (available in English, Spanish, French, Russian, Chinese and Arab). This website is larger than the OEMP, and it offers as well great quality of content. Thus, considering it as a potential candidate for further enhancements to the engine, we later used its English version to retrain the language model.

First, we crawled the OEMP version in Spanish[43], and after 5 hours, we then crawled the English version[44]. It was necessary to convert the html files into plain text. To do this, we ran the application named html2txt developed by Bobsoft.com[45]. The html files have been successfully converted to .txt in two clicks and have been automatically encoded to UTF-8. Since the scope of this crawling was to retrieve as much body text as possible. It was not

[41] http://www.oepm.es/es/index.html
[42] http://www.wipo.int/tools/es/sitemap.html
[43] https://www.oepm.es/es/index.html
[44] https://www.oepm.es/en/index.html
[45] This program is currently unavailable, but you can access the application through the following sharable link. Just click here.

necessary to run any boiler pipe scripts to remove tags, headers and footers, instead we continued with the alignment task in *LiveDocs*. In this case, the alignment function worked perfectly fine in most parallel sites. In most documents, the headers and footers were disregarded. See Figure 17. Only the body text was exported to the translation memory. The alignment of over 200 parallel texts was a very time-consuming task.

Figure 17. Alignment of html files from the OEMP. Only the confirmed links in blue were exported as a TM.



47.	Resultado de la investigación correspondiente a la denominación propuesta.	Search results for the proposed name.	47.
48.	Columna 6:	Column 6:	48.
49.	Clases	Classes	49.
50.	Indica la clase o clases del expediente.	Indicating the file class(es).	50.
51.	Corresponde a los códigos de la Clasificación Internacional de Productos y Servicios.	This refers to code of the International Classification of Goods and Services.	51.
52.	Informe de Parecidos	Similarity Report	52.
53.	La interpretación es la misma que para el caso de las Identidades con la excepción de que no aparece la columna de Clase ya que la búsqueda se hace en una sola clase.	The layout of the report is the same as the identity report, except that it does not include Column C since the search is performed on just one class at a time.	53.
54.	Ayudas y subvenciones Perfil del contratante Otras informaciones Estadísticas Centros Regionales de Información de PI Empleo Aula de Propiedad Industrial Internacionalización Transferencia de Tecnología Enlaces y direcciones de interés Recursos apoyo y prevención ante violencia género (WRAP) Portales OEPM Portal de la transparencia Gobierno de España Calidad Stopfalsificaciones Web Histórica CEVIPYME Campañas de sensibilización SÍGUENOS EN:	Grants and subsidies Contractor Profile Other Information Statistics Industrial Property Regional Information Centres Employment Industrial Property Education Internationalisation Technology Transfer Interesting addresses and links Recursos apoyo y prevención ante violencia género (WRAP) OEPM Portals Transparency Portal Government of Spain Quality Stopfalsificaciones Historic Web CEVIPYME Awareness campaigns FOLLOW US:	54.
55.	Suscripci&oacute;n a informaci&oacute;n Darme de alta Aviso Legal Accesibilidad	Suscribe to our Newsletter Register Legal notice Accessibility	55.

By the end of the collection of the Costa Rican corpus and the OEMP site, memoQ reports a total of 171,886 entries. The alignment project was exported into a single TMX (56,2 MB) and was saved as "ALIGNED_es_en," encoded in UTF8. See Table 10 for more details on the aligned documents.

Table 10. Additional Set of TMX Files Ready to train the Translation Model of System 2

Corpus name	Source/Translator
CR Electoral Code	Gabriela Castro
CR Political Constitution	ConstituteProject.org
2 CR Resolutions	Floria Sáez Rodríguez
OEMP Spain	Crawling Technique

4.1.4.3 World Intellectual Property Organization (WIPO)

Different from 4.1, we will not use the target side of the TMX for the monolingual texts for System 2; instead, we proceeded with another crawling technique to download the WIPO English website. We used the online SAAS platform named Sketch Engine, as it features an integrated boilerpipe and advanced html tag removal functions.

Sketch Engine uses the WebBootCat feature to download an entire site. After entering a new name for the corpus and selecting the target language, the cloud-based software asks for either seeds, urls (for specific paths) or a website link. In this case, we were interested only in the patents, English version of the site, so the following url was entered http://www.wipo.int/pct/en/. Additionally, this tool features advanced functions that allows the user to restrict the amount of cleaning, which "involves converting to plain text, removing boilerpipe and consolidating white spaces." In this case, the default settings were applied. After clicking on next, the crawling process begins.

This process can be performed offline, and once the harvesting process is over, it is necessary to compile the documents in one corpus. For this, Sketch Engine offers a series of options that will allow to compile a much cleaner corpus. For instance, one can select the type of sketch grammar, the removal of duplicates containing the "p"- paragraph tags and the compilation of a corpus containing only the "s"-sentences and "p"-paragraph tags. All these options were selected as it met the needs of the project. Once the corpus was successfully compiled, it was downloaded in plain text.

An additional Regex Clean-up process took place using Notepad++. Even though Sketch Engine applies its own cleansers based on the English boundaries set by the Sketch grammars, the corpus features a great amount of noise such as invalid broken sentences and numbers that break the segmentation of the text. Thus, the following Table 11 aims to summarize all the regex executed.

Table 11. Regex Clean-Up Summary for the WIPO Files

Find	REPLACE	FUNCTION
^([0-9]{2})$		To delete page numbers of two digits.
página [0-9]* de [0-9]*		To delete page numbers with the following format: "Página 01 de 10."
https?:\/\/(www\.)?[-a-za-z0-9@:%._\+~#=]{2,256}\.[a-z]{2,6}\b([-a-za-z0-9@:%_\+.~#?&//=]*)		To remove any URL.

<[^>]*>		To remove all html tags[46].
\n\d[^\n]*		To remove segments that begin with numbers.
\s+?$		To remove trailing whitespaces.
,\n^ |;\n^	,	To delete incorrect break lines after commas.
		To delete double spaces.
([\n\r]+)		To delete invalid line breaks and return carriages within sentences.
^		To delete blank spaces beginning
^$		To delete empty paragraphs.
^*		To delete asterisk at the beginning of strings
^\-		To delete dashes at the beginning of strings
--- * --1		To remove information of this kind.
^page ([0-9]*)$		To delete page numbers
[continued on next page]		To remove entire phrase

parent_folder="wipo"
id="file([0-9]*)"
filename=".*.pdf">
		To delete other kinds of urls with that structure

^([A-Z]{3})$ ^([A-Z]{2})$		To delete country initials at the beginning of strings such as USA, CR
^\d*$		To delete numbers at the beginning of strings
^\d\.\d$ ^\d\d\.\d$		To delete numbers like 3.3 and 54.2 at the beginning of strings
^\d\d\,\d\d\d$ ^\d\,\d\d\d$ ^\d\d\d\,\d\d\d$		To delete amounts of money at the beginning of strings such as 29,463, 5,054, 2,3900
^\(\d\d\,\d\d\d\)$ ^\(\d\,\d\d\d\)$ ^\d\d,\d\d\d\)$		To delete amount of money following this format: (12,535), (9,402), 25,680).
^\d\dbis.\d$		To delete numbers like 89bis.3
^<p> </p>$		To delete paragraph tags
^\d\d\.\d\d$		To delete numbers like 43.10
^\d\d\.\dbis$		To delete numbers like 66.1bis
^\d\d\.\dter$		To delete numbers like 66.1ter
^[A-Z]{3}*$		To delete currencies
^\(from \d\.\d\.\d\d\:$		(from 1.4.18:
⍰		To delete these signs ⍰
•		To delete these signs •
^X$		X at the beginning of strings

For more difficult cases of invalid segmentation of particular sentences, a series of find a replace searches were ran in Microsoft Word. First, we removed all paragraph marks, and

[46] A few of these regular expression were taken from the work of Peña, V. Entrenament de motors de traducció automàtica estadísticaentre el castellà i el romanésespecialitzats en farmàcia i medicina.

then all strings ending in “. “ were replaced by a “. [Paragraph mark]” After this, we have compiled a text of 3062 strings and saved it again under the name “source.utf8.trg.mono” file.

4.1.4.4 Glossary, Run-Time Glossary and Ignore List

The main distinction between a training glossary and a run-time glossary lies on their purpose. A run-time glossary is added to the Training tab, and it usually contains unique translations for terms. Run-time glossaries are very powerful search-and-replace tools and can be modified anytime without the need of retraining the engine. Training glossaries, on the other hand, are added to the Training tab, and they are used for multiple translations of a word, like varying forms of verbs and adjectives.

For the design of the run-time glossary, we used the translation of over 300 Costa Rican institutions, provided by TraduRed[47]. A series of corrections were made; for example, in case there were two possible translations for a term, only one term per source term remained.

Besides the unknown words reported from the first translation job of unseen strings (See Appendix 10. Unknown words that System 1 Fails to Recognize), we also looked for untranslated terms in the translation output text (See Appendix 11). A few of the words added to the run-time glossary were: “CrearEmpresa”, “sociedad en comandita”, “fundación”. See a list of examples[48] of the entries added to the run-time glossary in Appendix 12.

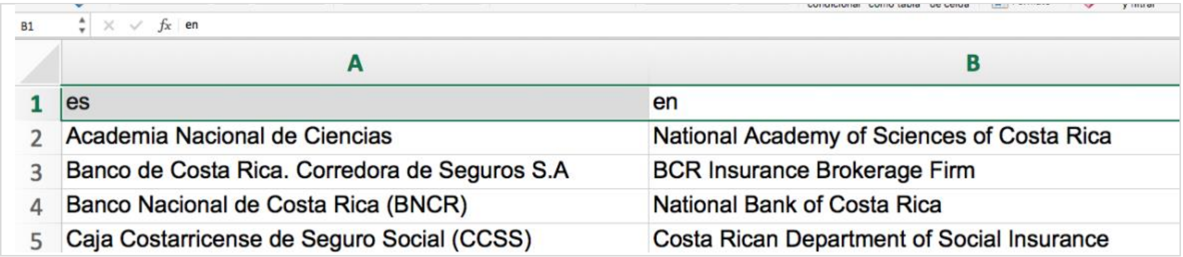
Likewise, the output text of System 1 (Appendix 11) served as the foundation to define the less specific terms to be added to the training glossary. An example of the added entries, include: “en línea” and “en línea con” respectively translated as “online” and “in line with”. Both entries are acceptable translations and can be present in future translation jobs. See a list of examples of the entries added to the training glossary in Appendix 13.

[47] https://tradured.com

[48] Click on the sharable link with the complete datasets and resources used for this methodology: https://drive.google.com/open?id=1PM_URmI_jeVa75mafttUNV4G-9N6J68B

Both glossaries were saved as “glossary.xlsx.” and in .xlsx format. Moreover, according to KantanMT’s specifications, it is crucial to add the source and target languages on top in the first row of the document, with the corresponding source and target languages (A1: {es} and A2: {en}), like illustrated in Figure 18. Failing to do this, KantanMT would interpret the files as translation memories.

Figure 18. Glossaries should be saved in .xlsx format.



	A	B
1	es	en
2	Academia Nacional de Ciencias	National Academy of Sciences of Costa Rica
3	Banco de Costa Rica. Corredora de Seguros S.A	BCR Insurance Brokerage Firm
4	Banco Nacional de Costa Rica (BNCR)	National Bank of Costa Rica
5	Caja Costarricense de Seguro Social (CCSS)	Costa Rican Department of Social Insurance

Regarding the ignore list file, we went over the 138 segments containing unknown words reported by the GapAnalysis as unknown words within the System 1’s phrase table. Many of these terms were false positives. Thus, we added the false positive terms in an ignorelist.txt, UTF-8 encoded file. The file was then added to the training tab for the retraining of System 1. The rest of the terms were either added to the run-time glossary or training glossary. By the end the refinement phase, set below is a list of the new training datasets, ready to retrain System 1.

Table 12. Training Dataset II

Name	Size
DGT_es_en_01.tmx	122 MB
DGT_es_en_02.tmx	148 MB
DGT_es_en_03.tmx	139 MB
DGT_es_en_04.tmx	145 MB
DGT_es_en_05.tmx	133 MB
DGT_es_en_06.tmx	31 MB
ECB_es_en.tmx	46 MB
News_es_en.tmx	91 MB
General 2 (Source WC: 6,851,405, Target WC: 7,318,088)	
ALINGNED_es_en.tmx	56,2 MB
source.utf8.trg.mono	409 MB
Run-time glossary	33 KB
Training glossary	27 KB
ignorewordslist	559 bytes

4.2 Training a Neural MT in Modern Machine Translation

Complying with one of the objectives of this research paper, we attempted to train a neural MT to determine which data-driven approach would be more suitable for our specific use-case. Nonetheless, after a series of tests, we concluded that current NMT platforms, particularly MMT and Kantan Neural MT, are still far from the reach of an average translator. See Appendix 14 for a complete account on the preparation and training attempts using MMT. See Appendix 15 for an account on using the datasets collected from section 4.1 on.

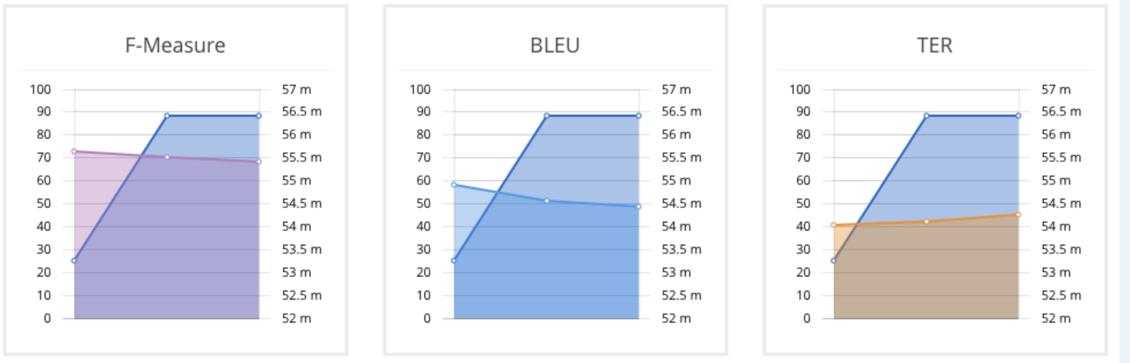
5. Analysis and Results

To determine which system is more proficient, we will carry out a comparative evaluation of System 1 and System 2 using again Build Analytics, KantanAnalytics, and additionally Kantan LQR integrated function. In section 5.2, through a manual, qualitative, human-based evaluation, we will discuss the most relevant types of errors found per engine and determine which engine produces the most desirable output text for post-editing purposes.

5.1 Re-measuring System 1 and System 2

Based on Figure 19, the refinement steps have resulted in a significant loss of quality. In fact, opposite to the expected results, all automated metrics dropped. In the graphs from Figure 19, the blue lines indicate the positive increase of training data, the other lines in the back represent the decrease of the F-Measure, BLEU and TER scores, accordingly.

Figure 19. A Summary of System 2's Quality Metrics Results

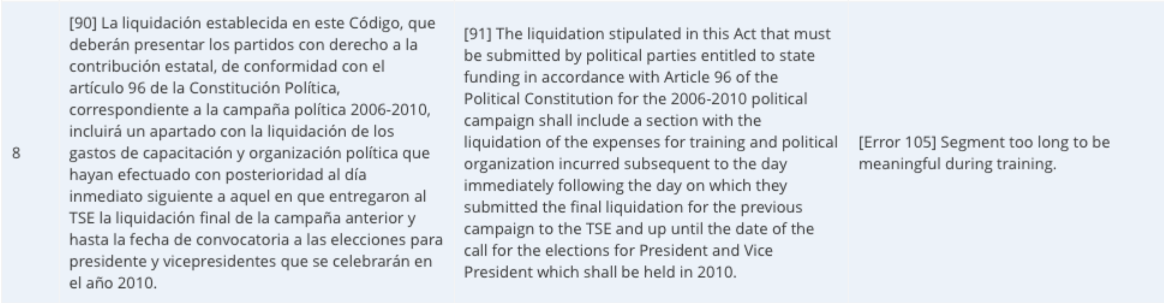


	System 1	System 2	Difference
F-Measure	70%	68%	-2%
BLEU	57%	49%	-8%
TER	43%	45%	-2%
WC	53,246,174	56,406,825	+ 3,160,651
Mono WC	31,935,856	49,672	- 31,886,184

Upon this, we proceeded with an inspection about how the new datasets and glossaries have affected the performance of System 2. The latest system counts with more than 3 million words in relation to System 1. According to the Training Candidates Rejects Report, the most recently added translation memory (ALIGNED.tmx) has not affected the overall percentage of rejected words and segments. Even though a total of 5% (151.960 words) were rejected, the average percentage of rejected words remained the same, 9%. See Appendix 17 for a complete report of the rejected words per translation memory.

It is important to note that most of the rejected segments reported correspond to false positives. Although the alignments were made correctly, still Kantan Data Cleansers lack of an understanding of the alignments of long segments, characteristic of legal corpora such as the Electoral Code of Costa Rica or the DGT corpora or KantanMT might feature a word number limit per segment, which should be considered prior to performing the preparation phase. Alternatively, a different aligning approach should be taken for future data preparation and training processes such as removing placeholders prior to aligning.

Figure 20. Segments of the Aligned, TMX file are Rejected due to Error 105



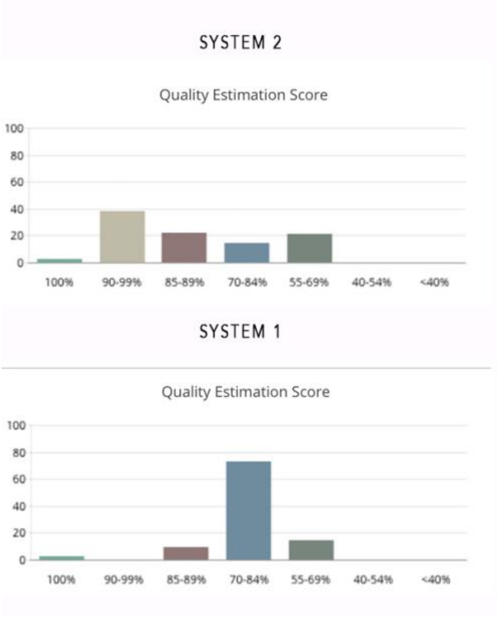
8	[90] La liquidación establecida en este Código, que deberán presentar los partidos con derecho a la contribución estatal, de conformidad con el artículo 96 de la Constitución Política, correspondiente a la campaña política 2006-2010, incluirá un apartado con la liquidación de los gastos de capacitación y organización política que hayan efectuado con posterioridad al día inmediato siguiente a aquel en que entregaron al TSE la liquidación final de la campaña anterior y hasta la fecha de convocatoria a las elecciones para presidente y vicepresidentes que se celebrarán en el año 2010.	[91] The liquidation stipulated in this Act that must be submitted by political parties entitled to state funding in accordance with Article 96 of the Political Constitution for the 2006-2010 political campaign shall include a section with the liquidation of the expenses for training and political organization incurred subsequent to the day immediately following the day on which they submitted the final liquidation for the previous campaign to the TSE and up until the date of the call for the elections for President and Vice President which shall be held in 2010.	[Error 105] Segment too long to be meaningful during training.

Additionally, we ran again a GapAnalysis[49] to confirm that the ignore word list has worked. Although many of the words added to the ignore list did not show up, a few of the reported false positives remained, like “India,” “idea,” “no” and “central.” In spite of the effort to detect the false positives during the first refinement phase, this makes evident the lack of control that the translator has over the engine’s output.

For further analysis on the impact that the creation of the general and run-time glossaries might have had in System 2, we ran a Translation Analysis. The analysis reports a substantial improvement of over 50% of Kantan Total Recall, which overall suggests a significant reduction in post-editing costs. See Figure 21 for further reference.

[49] Access to the complete list of unknown words https://drive.google.com/open?id=1VujGkyEkFQAvH4KXTtwhDF8xws438rC3

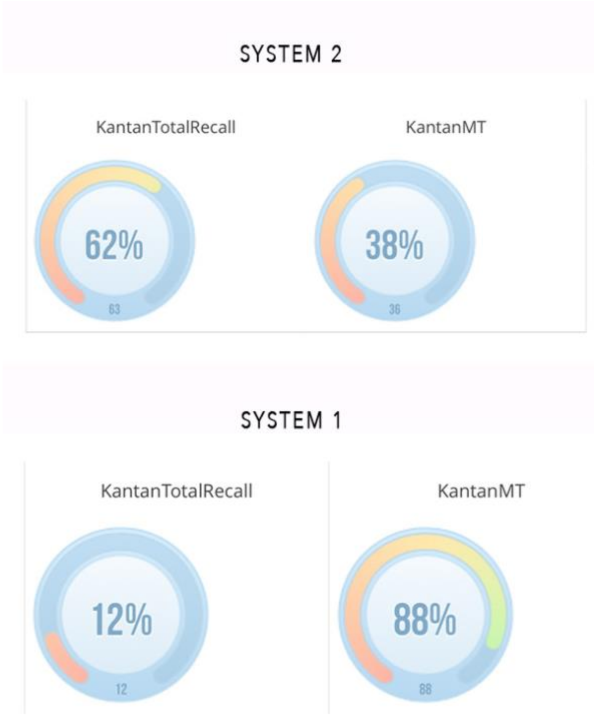
Figure 21. Translation Analysis Cost of System 2 and System 1



The KantanAnalytics Report provides match statistics both, at a word and segment level. Based on Figure 21, it is evident that System 2 has now a much richer vocabulary than System 1 thanks to the glossaries provided. System 2 features 38,1% of 90-99% word matches, while System 1 only has a 0,4% of 90-99% word matches. Although a few words, especially verbs, had been included in the glossary, they are still identified as unknown by System 1. It might be necessary to take a rule-based approach, rather than a data-driven focus.

At a segment level, KantanAnalytics reports that System 1 offers 58% more translation capacity than System 2. See Figure 22 for exact figures. However, sometimes the relevance of the metrics provided by KantanMT can be misleading since based on these same reports, both engines have the capacity to translate 75 out 76 segments found in the Source Testing Text. The dilemma, then lies on the actual usability of that output text.

Figure 22. Comparative of KantanAnalytics. Report on Total Recall and MT Capacity to Produce Translations



When it comes to comparing two or more engine's performance, the measuring reports can be misrepresentative of what the actual quality of the output text might be. Therefore, in the following section, we exited the KantanMT environment and proceeded with a different evaluation model that would help to determine which engine's output is more fulfilling (has a higher "passing" threshold value).

5.2 Linguistic, human-based evaluations of System 1, System 2 and Google Translate

Using the "Profile Your Content" site, TAUS DQF recommends to apply the Error Typology Model based on the project's scope and type of content[50]. See in Appendix 18 the selected options. The Error Typology Evaluation [51] is a manual, human-based evaluation that enables

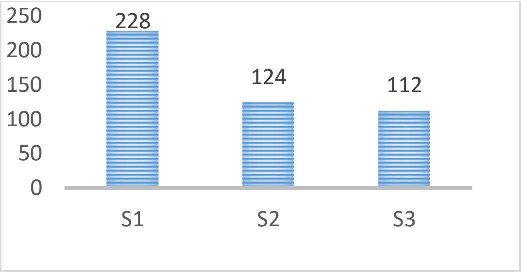
[50] Additionally, a ranking evaluation to assess the engines' output quality was performed with the aid of two translators. Since results correlate quite well with the Error Typology Model, only the latter is further explained in this chapter. Otherwise, see Appendix 20 for more details on the Ranking Evaluation Results.

[51] According to the results of the "Profile Your Content from TAUS," TAUS suggests to use this evaluation model: https://www.taus.net/knowledgebase/index.php?title=Error_typology

companies or translators to run a consistent, yet customizable assessment. It works with small sample sets and the user can define the criteria and severity of the errors.

For this evaluation, the output texts of System 1, System 2 and Google Translate (S3 from now on) went under examination. It was decided to evaluate Google's output since it is one of the most popular choices of MT engines among language providers and end-users[52]. Moreover, regarding the evaluation's settings, the reviewer[53] (in this case, the researcher of this project) looked up for errors of fluency, adequacy, accuracy, terminology and style. We will first discuss the global results, and then the most relevant cases of penalized errors.

Figure 23. Global Results of the Error Typology Test



According to the global results, S1 ranks as the first worse engine, with a total of 228 errors in a sample test of 48 segments and 711 words. In addition, Google Translate ranks as the best engine among this group with 112 errors, while S2 ranks second best with a total of 124 errors. This graphic not only makes evident that S2 has effectively improved after the refinement steps, but also reinforces what has been concluded in 5.1, automated metrics indeed have misrepresented the actual expected quality of the output text.

It is noteworthy to highlight the engines' strengths and limitations this evaluation reports. This human-based evaluation helps us to get a deeper understanding of the inner workings of the machine translators. The analysis is structured based on the selected Error Typology Classification.

[52] Based on a study from 2016, Google translates 143, 280, 496, 726 words per day. For more information, refer to Way, A. *Quality Expectations of Machine Translation*.

[53] Although the evaluation of a MT engine should be performed by a third person, due to its high cost, it was not possible to find a candidate different from the researcher to perform this task. The full description of each of the errors to be found and their specific criteria was intended to protect the project from any scientific risk to bias.

5.2.1 Fluency Errors

The fluency category focuses on grammar, spelling, correct use of titles and names and its readability for a native speaker. The following 4-point scale of fluency provided by TAUS was used:

Table 13. The 4-point Adequacy Rating Scale Defined by TAUS
4 Č Flawless refers to a perfectly flowing text with no errors. 3 Č Good refers to a smoothly flowing text even when a number of minor errors are present. 2 Č Disfluent refers to a text that is poorly written and difficult to understand. 1 Č Incomprehensible refers to a very poorly written text that is impossible to understand.

In addition, it was resolved to automatically penalize the terminology criteria whenever an error for the mistranslation of proper names applied. The decision was founded on the fact that the expected terminology has been carefully curated during the refinement stage, thus a crucial requisite for the engine was to comply with the glossary specifications given.

Table 14. Most Common Fluency Errors

ID	Source	Output from S1.	Golden Standard Translation
12	g. Certificado de uso de suelo.	g. Land use certificate.	g. Certificate of land use.
		Output from S2.	
6	a. Inscripción de sociedades mercantiles en el Registro Nacional.	a. Registration of Corporations in the National Register.	a. Registration of corporations in the National Registry.
		Output from S2.	
21	La inscripción de una empresa es la creación jurídica de una nueva sociedad ante el Registro Nacional. Este trámite lo puede realizar únicamente el notario.	The registration of a company EN legal the creation of a new company with the national register. This procedure can be carried out only the notary.	The registration of a company refers to the creation of a new and legal society within the National Registry. Only notaries have the legal authorization to complete this procedure.

For example, in segment 12, the output from S1 was rated as flawless, but one point was marked down since the correct term for “certificado de uso de suelo” had been added into the run-time glossary. This penalization does not affect the fluency rate, though. In the case of segment 6, there are two similar examples in which the term and phrase “sociedades mercantiles” and “Registro National” were previously translated as “corporations” and “National Registry.” Here, three errors were penalized. Besides the two terminology errors, the wrong capitalization of “Corporations” affected the fluency score. In the case of segment 21, S2 struggles with grammar structures, like Verb+to be in simple present, which is a critical error that propagates throughout the output text, affecting the fluency and adequacy of the translation.

5.2.2 Adequacy Errors

Adequacy focuses on how much of the translation conveys the original message. The 4-point scale of adequacy provided by TAUS was also used.

Table 15. The 4-point Adequacy Rating Scale Defined by TAUS
4 Č Everything. All the meaning in the source is contained in the translation, no more, no less. 3 Č Most. Almost all the meaning in the source is contained in the translation. 2 Č Little. Fragments of the meaning in the source are contained in the translation. 1 Č None. None of the meaning in the source is contained in the translation.

Likewise the criteria to penalize fluency errors, adequacy points were taken whenever there were terminology or style errors. See Table 16.

Table 16. Common Adequacy Errors

ID	Source	Output from S3.	Golden Standard Translation
40	Toda información debe ser corroborada en La Gaceta y el Alcance correspondiente.	Information should be supported in the Gazette and the scope for compensation.	All information must be supported in the corresponding official journal La Gaceta and Alcance.

Despite of the fact that S3 does not correspond to a trained engine, errors of terminology and style also applied. “La Gaceta” and “el Alcance” have been added to the run-time glossary, but as illustrated in segment 40, S3 is not able to reflect the preferred terminology and/or style. Besides, note that S3 fails to identify the name of a newspaper, seriously affecting the translation.

5.2.3 Accuracy Errors

An accuracy error is closely related to fluency and adequacy. It takes into account additions, omissions and mistranslations. During the evaluation, it was determined to penalize the accuracy rate, whenever a fluency or adequacy error was detected, depending on the case. Segment 31 exemplifies a case of addition, although it does not affect the adequacy score. Segment 7 refers to a case of omission; in this instance, it has led to a loss of meaning.

Table 17. Common Accuracy Errors

ID	Source	Output from S1.	Golden Standard Translation
31	Hace la solicitud de conformidad con el artículo 88, inciso 2) de la Ley General de Migración y Extranjería 8764, vigente a partir del 01 de marzo de 2010, cumpliendo con todos los requisitos.	Makes the application in accordance with Article 88, point 2)the General law on Migration and Foreigners N-8764, in force from 01 March 2010, compliance with all the requirements.	Makes the request in accordance with Article 88, paragraph 2) of the General Law on Migration and Aliens 8764, effective as of March 1, 2010, complying with all requirements.
		Output from S2.	
7	La compañía (nombre formal de la compañía), solicita autorización de estancia como visitante de negocios a favor del Señor(a) (nombre), mayor, (estado civil), Agente de Negocios, vecino de (dirección en país de origen), de paso por (dirección en Costa Rica), Costa Rica, ciudadano (nacionalidad), … compras públicas Mer-link.	The company (formal name of the carrier), applying for residence permit as business guest Mr (a) in favour of (name), of legal age, (marital status), staff member, turnover, resident of (address in the country of origin), crossing (address in Costa Rica), Costa Rica, citizen (nationality), …	The company [official name of the company] requests a residence permit as a business guest, on behalf of Mr. or Mrs. [name], of legal age, [marital status], business agent, resident of [address from its country of origen], temporarily residing in [address in Costa Rica], citizen [nationality], ….

5.2.4 Terminology

The adherence to terminology of the public sector of Costa Rica was strictly evaluated. This aspect also included another sub-category named local conventions, which in the original DQF Evaluation is deemed as a separate error type. Since we had added a run-time glossary and a general glossary where the US spelling style is the most preferred in Latin-American countries (using a "z," instead of "s" in nouns such as "legalization"), failing to follow that specific spelling preference was considered as an error of terminology as well as style. The incorrect translation of procedures or descriptions would slightly affect the usability of the output text.

Table 18. Common Terminology Errors

ID	Source	Output from S3.	Golden Standard Translation
8	c. Publicación de edictos en el diario oficial, La Gaceta.	Publication of edicts in the official gazette, La Gaceta.	c. Publication of edicts in the official journal La Gaceta.
		Output from S1.	
10	f. Certificación de planos catastrados digitalizados.	f. Certification of planes catastrados digitised.	f. Certificate of digitized cadastral plans.

5.2.5 Style

The use of a consistent style was highly valued during the examination. Any issues with adequacy would in many instances affect the style category. This aspect values awkwardness, inconsistent style, third-party style and unidiomatic structures. When referring to style, in a real-life scenario, the reviewers of the MT will have to support their decision upon their clients' style guide; in this case, the reviewer should make reference to the golden standard translation, which can be found in Appendix 19.

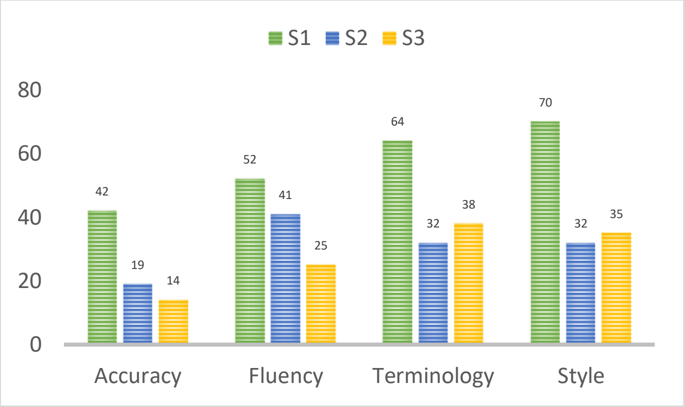
Table 19. Common Style Errors

ID	Source	Output from S3.	Golden Standard Translation
20	5. ¿Qué es la inscripción de una empresa?	5. What is the registration of a company?	5. What does registering a business mean?
		Output from S1.	
10	f. Certificación de planos catastrados digitalizados.	f. Certification of planes catastrados digitised.	f. Certificate of digitized cadastral plans.

As it was just explained, segment 10 illustrates the wrong use of the US spelling style, and segment 20 depicts an issue in which the translation was done too literal, conveying little meaning to the target user.

Having set the grounds of evaluation, the following are the general results based on the number of errors found in the output texts of S1, S2 and S3. See Figure 24.

Figure 24. Summary of the Error Typology Results



Based on the final results of the Error Typology Evaluation, S3 is far superior than S2 and S1, in terms of accuracy and fluency. While in terms of terminology and style, S2 beats S1 with an advantage of 6 and 3 points. Regarding accuracy, the advantage of S3 over S2 is of only 5 points. Fluency, then, seems to be the main reason why S3 stands out from S2. It is worth to point out that there was a significant improvement from S1 to S2 in all criteria, except Fluency (a gap of 11 points). In terms of Accuracy, S2 beats S1 with a great advantage of 23 points and Terminology and Style with an edge of over 30 points. This shows that from a human judgement, data-driven refinements actually affect measurable, positive results.

Regarding the possibilities and expectations of S2 and S3, either output could be potentially used for assimilation or post-editing purposes. While engine S3 produces more adequate and fluent translations, using S2 could reduce the translation cost related to the correction of terminology. On the other hand, different workarounds could be employed in order to speed up the post-editing of S3 such as integrating the glossaries into a CAT environment. At this stage of the research, it would be necessary to study further what process takes longer (correcting terminology or fluency) in order to determine which engine proves more beneficial if deployed.

6. Conclusions

Every step followed in this experiment was an opportunity to learn. Thus, below are a few clarifications regarding some of the procedures registered in the methodology that should be avoided. Also, a few recommendations are proposed for further development of this project.

Regarding the **data preparation phase,** it was later discovered that the step of inverting TMX files in accordance to the engine's language combination is not necessary, so that step should be completely dismissed. In addition, when **preparing the datasets,** it would be important to add a regex to remove the placeholders. That would help to reduce the 9% of rejected words of the training data, taking more advantage of the resources that the researcher has already worked hard to collect and prepare.

Also, when **collecting data**, it is strongly recommended to narrow down as much as possible the domain. As seen in this research, the e-government domain is very ample, but at the same time narrow because most websites' content overlap, in terms of sections (most websites include regulations, policies, FAQ, general help guides, etc.). In this case, a suggestion would be that a future, more robust e-government encompasses a suite of engines sort by sectors such as health, education, immigration, finances, legal and tourism. In this way, the choice of engine would depend on which section or content of the website requires translation.

According to the **measuring and assessment** results, it is evident that the automated metrics should remain as a quick review of the possible training outcomes (except for the KantanTotalRecall Rates and Quality Estimation Score). As shown in this research, although more expensive and time-consuming, human-based evaluations correlate better with the translator expectations of quality. In this research, only two human-based evaluations were carried out due to the scope of this research, but an additional evaluation regarding its productivity rate should be carried out, as explained at the end of the Analysis and Results.

Regarding the **refinement stage**, it is clear that Sketch Engine is far superior than HTTrack Website Copier. Although the latter one is for free, Sketch Engine features more advanced extraction algorithms, resulting into cleaner web crawls. In addition, it is important to highlight that the use of other approaches besides the corpus-based ones would be necessary to improve fluency of S2. As demonstrated in 5.2.1Fluency Errors, the output errors of S2 can be easily resolved through the creation of a rule for simple present conjugations; i.e., rule-based approaches.

Regarding the choice of **the training environment**, KantanMT is well-equipped to refine and assess PB-SMTs. However, it is crucial to highlight the fact that S2 has not reached its full potential by the end of this research project. Further improvements (adding post-edited corrections, rule-based approaches) should be carried out to raise a much more mature, fulfilling engine. Although this can only be achieved in the long-term, PB-SMT technologies again prove to be superior as it fulfills the needs for customization. On the contrary, the use of Google Translate's output or any other generic MT's text cannot be manually refined or controlled, leading to counterproductive translation costs (correcting more than once the same term, unless it meets a fuzzy match). Moreover, at least the customization of PB-SMT systems give translators control over all stages of the production process, while NMT technologies, as shown in 4.2, are less inclusive at present[54] due to economical, knowledge-based factors. In spite of this last fact, NMT is still a very promising technology, and once it becomes more accessible for translators, researchers and translators would be able to benefit from the training of engines using deep learning technologies and high quality training data.

This research proves that using a corpus-based approach is feasible, but not exclusive. If any interest raises to develop the complete localization of the CR e-government, then this research should work as a starting point to train a more robust engine. As stated in the Introduction, this research has only contemplated the translatability of employing data-driven approaches such as using the most relevant open data resources in the combination ES_ES Č EN. To reach best levels of proficiency, it would be key to consider creating a style guide for these institutions and pre-editing the source content to obtain a more controlled language. Another important finding from this research is the fact that any Latin American,

[54] On May 7th, 2018, Microsoft announced the launch of custom NMT functionalities named Custom Translator.

Spanish speaking country could benefit from this technology to improve their digital government's content as well. The most important aspect to be considered is to carefully collect and curate the public sector terminology, which differs in every country. To pursue that is not simple, as many procedures or names of institutions were provided as "non-official," but just as a reference.

In a continuous effort of affecting positive change (the acceptance and use of MT technology on behalf of translators), a video tutorial series named "MT for Everyone: how-to train a corpus-driven MT" was produced with the support of Qabiria. Along this, given the fact that preparing datasets can be the bottleneck to pursue research in the combination of ES_CR Č EN, a public repository is now available under the name "Open-Data-for-MT_ES-CR-EN[55]." This project is meant to support more research and to facilitate more resources to translators interested in MT systems. Today, it counts only with the three datasets that were used for the training of S2, but future additions are already in progress. Finally, as a possible next step towards advancement on this field in CR would be to enforce a national Open Data Policy for translation, so that machine translation efforts (such as this research project) can bring about development in the academia sphere and the public and private sectors.

[55] https://github.com/gcocozza/Open-Data-ES-CR-EN-/tree/master

7. Bibliography

Adequacy/Fluency Guidelines. (2013). In TAUS Enabling Better Translation. Retrieved on 04/16/2018, from https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines

Agenda Digital de España. (n.d.). *Plan de Impulso de Tecnologías del Lenguaje (Plan TL). Presentación del plan.* Retrieved from http://www.agendadigital.gob.es/tecnologias-lenguaje/Paginas/plan-impulso-tecnologias-lenguaje.aspx

Aragonés, M. & Way, A. (2017). On the Complementarity between Human Translators and Machine Translation. *Hermes-Journal of Language and Communication in Business, 56*, 22-41. http://dx.doi.org/10.7146/hjlcb.v0i56.97200

Arias, P. (2017, August 10). Costa Rica acaparó el 27% de la inversión extranjera en Centroamérica 2016. *CRhoy.com*. Retrieved from https://www.crhoy.com/economia/costa-rica-acaparo-el-27-de-la-inversion-extranjera-en-centroamerica-en-2016/

Barahona, JC. & Elizondo, A.M. (2010). Evaluación de Sitios Web del Gobierno y Municipalidades de Costa Rica 2010. *INCAE, Business School*. Retrieved from http://www.incae.edu/images/descargables/Noticias/INFORME_2010.pdf

Britz, D. (2016, January 3). Attention and Memory in Deep Learning and NLP. Retrieved from http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/

Chung, J. (n.d.). *5.2 Deep Learning with RNN (Recurrent Neural Network)* [Video]. Retrieved from the Coursera website: https://www.coursera.org/learn/deep-learning-business/lecture/AuemB/5-2-deep-learning-with-rnn-recurrent-neural-network

Cattelan, A. (n.d.). Phrase-Based Vs Neural MT – Webinar Questions [Video]. Matecat. Retrived from https://www.matecat.com/phrase-based-vs-neural-mt-webinar-questions/

Dolz, J. (2016). *Competencias profesionales asociadas a la traducción automática estadística.* Retrieved from Repositori Universitat Jaume I Dissertation. (TI0983)

European Commission. (2011). COMMISSION DECISION of 12 December 2011 on the reuse of Commission documents. *Official Journal of the European Union*. Retrieved from http://data.europa.eu/eli/dec/2011/833/oj

Faherty, L. (2017, October 25). *Kantan Analytics*. Retrieved from the KantanMT website: https://kantanmt.zendesk.com/hc/en-us/articles/205319985-KantanAnalytics-Video-

Farajian, Aming. M, Turchi, M. (2017). Multi-Domain Neural Machine Translation through Unsupervised Adaptation. *Proceedings of the Conference on Machine Translation (WMT)*, 1, 127-137. http://www.aclweb.org/anthology/W17-4713

Flores, B. 2018. (2018, May 14). Costa Rica es el primer país en cuantificar exportaciones de servicios vía TIC's. LaRepública. Net. Retrieved from https://www.larepublica.net/noticia/costa-rica-es-el-primer-pais-en-cuantificar-exportaciones-de-servicios-via-tic-s

Galbrun, E. (2009). *Phrase Table Pruning for Statistical Machine Translation*. Retrieved from Loria, Laboratoire Lorrain de Recherche en Informatique et ses applications https://members.loria.fr/EGalbrun/resources/Gal09_phrase.pdf

Germann U. (2015). Sampling Phrase Tables for the Moses Statistical Machine Translation System. *The Prague Bulletin of Mathematical Linguistics, 104, 39-50*. doi: 10.1515/pralin-2015-0012

Guerra, L. M. (2003). *Human Translation versus Machine Translation and Full Post-Editing of Raw Machine Translation Output.* Retrieved from the Dissertations database of the Dublin City University. http://sceuromix.com/enlaces/MASTER%20IN%20TRANSLATION%20STUDIES%20BY%20LORENA%20GUERRA-2003.pdf

Haddow, B. (2014, July 7). Principles of Machine Translation. TAUS Machine Translation and Moses Tutorial. [Video] Retrieved from https://www.youtube.com/watch?v=beX5rqdneII

Harris, I. (2004). *Floating Point Numbers.* Retrieved from https://www.doc.ic.ac.uk/~eedwards/compsys/float/

Hutchins, J. (Eds.). (2000). *The First Decades of Machine Translation. Early Years in Machine Translation.* John Benjamins Publishing Company. Amsterdam.

Hutchins, J. n.d. *The Practical use of MT systems.* Retrieved from http://www.hutchinsweb.me.uk/IntroMT-8.pdf

Jörg, T. 2012, *Parallel Data, Tools and Interfaces in OPUS.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)

Kalniŋš, R. n.d. *Machine Translation: Enabling Multilingual Communication in e-Government* [PDF Document]. Retrieved from the European Language Resource Coordination website: http://lr-coordination.eu/sites/default/files/Iceland/Kalnins_ELRC_Iceland%20FINAL.pdf

KantanMT. (2018, April 26). *How to Create a Run-time Glossary*. Retrieved from https://kantanmt.zendesk.com/hc/en-us/articles/115002707203-How-to-Create-a-Run-time-Glossary)

Kenny, D., Doherty, S. (2014). Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer,* 8:2, 276-294, doi: 10.1080/1750399X.2014.936112

Koehn, P. (n.d). Europarl: A Parallel Corpus for Statistical Machine Translation. *University of Edinburgh, Scotland.* Retrieved from http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

Koehn, P. (2013, March 25). Speech: Open Problems in Machine Translation [Video]. Retrieved from https://www.youtube.com/watch?v=6UVgFjJeFGY

Koehn, P. (2015). *Machine Translation and the Translator* [PDF Document]. Retrieved from DGT TRAD Terminology Coordination. http://www.termcoord.eu/wp-content/uploads/2015/04/03koehn.pdf

Koehn, P. & Knowles, R. (2017). Retrieved from John Hopkins University. *Six Challenges for Neural Machine Translation* [PDF Document]. Retrieved from http://www.aclweb.org/anthology/W17-3204

Lau, P. (1987). Eurotra: past, present and future. Commission of the European Communities. *Translating and the Computer, 9.* Retrieved from https://pdfs.semanticscholar.org/c63d/e759849272035a0c8cba9ea6a983b2b8658b.pdf

Machine Translation and Governments: The Case of Canada. (2016, April 26*). Morningside Translations.* Retrieved from https://www.morningtrans.com/machine-translation-and-governments-the-case-of-canada/

Machine translation for public administrations — eTranslation. (n.d.). *European Commission.* Retrieved from https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en#product-features

Marcello, F., Bertoldi, N., Caroselli, D., Cattoni, R., Cettolo, M., Germann, U., Mastrostefano, L. & Trombetti, M. (2016). *D3.1. MMT-Modern Machine Translation. First Report on Database and MT Infrastructure.*

Martín, A. & Piqué, R. (2017). MTradumàtica i la formació de traductors en Traducció Automàtica Estadística. *Revista Tradumática,* 15, 1-11. https://doi.org/10.5565/rev/tradumatica.199

Microsoft Translator. (2018). What Is a Neural Network Based Translation? Retrieved from https://translator.microsoft.com/help/articles/neural/

Mikolov, T., Sutskever, I. and Le,Q. (2013, August 14). *Learning the meaning behind words*. Retrieved from https://opensource.googleblog.com/2013/08/learning-meaning-behind-words.html

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean J. (n.d.). Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems Foundation, Inc.* Retrieved from https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

O'Dowd, Tony. (2017). *Get Started with KantanNeural* [PDF Document]. Retrieved from https://es.slideshare.net/kantanmt/get-started-with-kantanneural

Peña, V. (2017). *Training statistical machine translation engines in the pharmaceutical and medical domain between the Romanian and Spanish languages*. MA Thesis retrieved from Universidad Autónoma de Barcelona.

Pérez, M.I. (2016). *Phrase-Based Statistical Machine Translation. Explanation of its processes and statistical models and evaluation of the English to Spanish translations produced.* BA Thesis retrieved from Universidad de Alicante. (32524).

Uszkoreit, H. & Koehn, P. (2006). EuroMatrix: Statistical and Hybrid Machine Translation Between All European Languages. Information Society Technologies. Retrieved from http://www.euromatrix.net

Quoc V. Le & Mike Schuster. (2016, September 27). A Neural Network for Machine Translation, at Production Scale. Google AI Blog. Retrieved from https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Salas, L. D. (2014, July 6). Gobierno Digital de Costa Rica logró posición histórica. *El Financiero. Retrieved from* https://www.elfinancierocr.com/tecnologia/gobierno-digital-de-costa-rica-logro-posicion-historica/7DLBEANGJNFJ3EHX54DII6GHOM/story/

Khalilov, M. (2012, September 26). Evaluating MT Systems. TAUS Machine Translation and Moses Tutorial. Retrieved from https://www.youtube.com/w atch?v=_v1j0e0IfWs

Toral, A., Esplà, M., Klubička, F., Ljubešić, N., Papavassiliou, V., Prokopidos, P., Rubino, R. & Way, A. (2016). Crawl and Crowd to Bring Machine Translation to Under-resourced Languages. Translators, 5, 205-226. Retrieved from Springer: http://dx.doi.org/10.1007/s10579-016-9363-6

Way, A. & Hearne, M. (2011). On the Role of Translations in State-of-the-Art Statistical Machine Translation. *Language and Linguistics Compass*, *5*(5), 227–248. https://doi.org/10.1111/j.1749-818X.2011.00275.x

Way, A. & Hearne, M. (2011). Statistical Machine Translation: A Guide for Linguists and Translators. *Language and Linguistics Compass*, *5*(5), 205-226. doi:10.1111/j.1749-818X.2011.00274.x

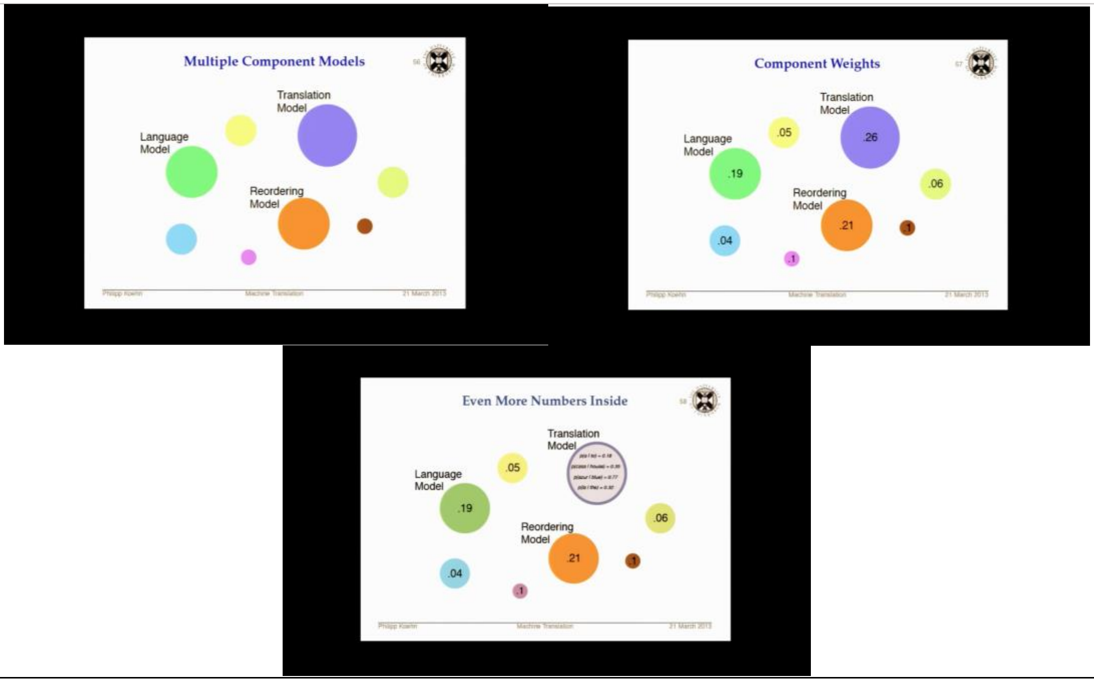
Way, A. n.d. Quality expectations of machine translation. Adapt Center, School of Computing, Dublin City University. Retrieved from https://arxiv.org/pdf/1803.08409.pdf

Zhaopeng, T. Zhengdong, L., Yang, L., Xiaohua, L. & Hang, L. (2016). Modeling Coverage for Neural Machine Translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.* Retrieved from http://www.aclweb.org/anthology/P16-1008

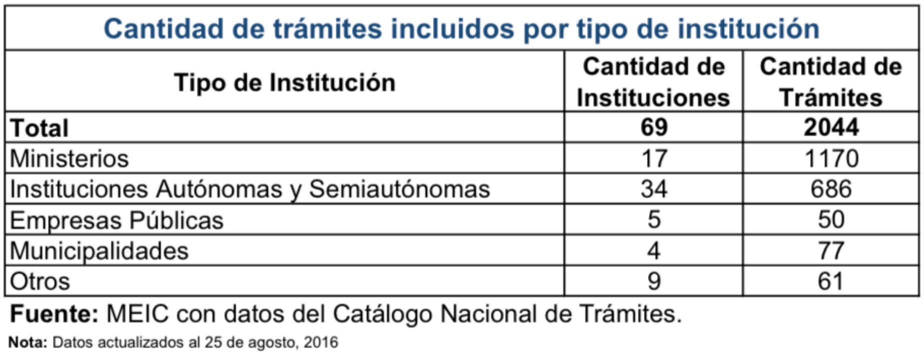
Wu, Y., Schuster, M., Chen, Z., Le, Q.V., & Norouzi, M. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. Retrieved from arXiv, 2. https://arxiv.org/pdf/1609.08144.pdf

8. Appendixes

Appendix 1. The distribution of weights of PB-SMT engines, according to Philipp Koehn.



Appendix 2. CR e-government statistics report



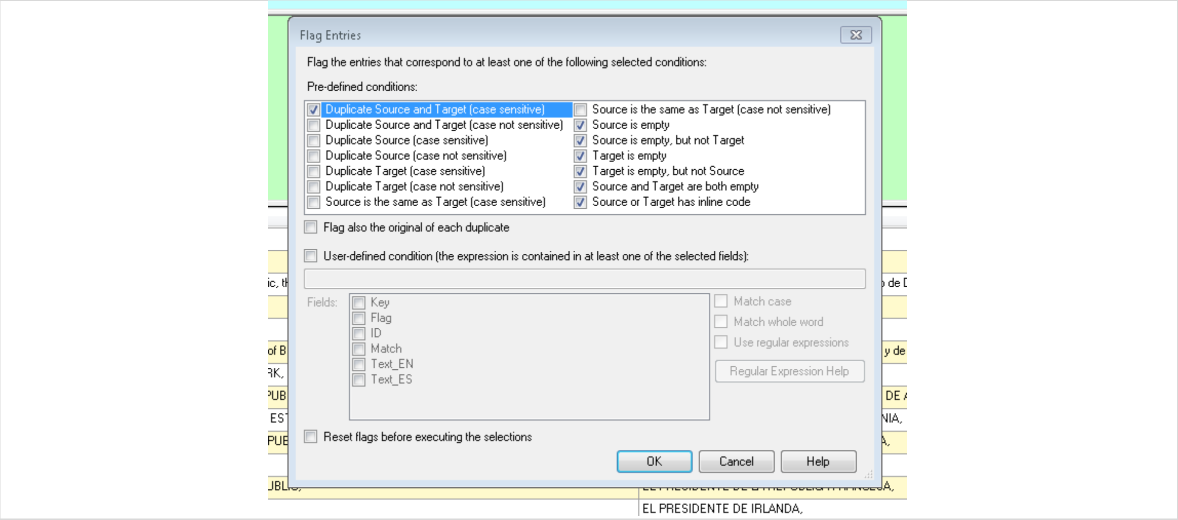
Cantidad de trámites incluidos por tipo de institución

Tipo de Institución	Cantidad de Instituciones	Cantidad de Trámites
Total	**69**	**2044**
Ministerios	17	1170
Instituciones Autónomas y Semiautónomas	34	686
Empresas Públicas	5	50
Municipalidades	4	77
Otros	9	61

Fuente: MEIC con datos del Catálogo Nacional de Trámites.

Nota: Datos actualizados al 25 de agosto, 2016

Appendix 3. Tasks involved in the Clean-up process of the Training Data. “Flag Entries” function found in Olifant.



Flag the entries that correspond to at least one of the following selected conditions:

Pre-defined conditions:

☑ Duplicate Source and Target (case sensitive)
□ Duplicate Source and Target (case not sensitive)
□ Duplicate Source (case sensitive)
□ Duplicate Source (case not sensitive)
□ Duplicate Target (case sensitive)
□ Duplicate Target (case not sensitive)
□ Source is the same as Target (case sensitive)
□ Source is the same as Target (case not sensitive)
☑ Source is empty
☑ Source is empty, but not Target
☑ Target is empty
☑ Target is empty, but not Source
☑ Source and Target are both empty
☑ Source or Target has inline code

□ Flag also the original of each duplicate

□ User-defined condition (the expression is contained in at least one of the selected fields):

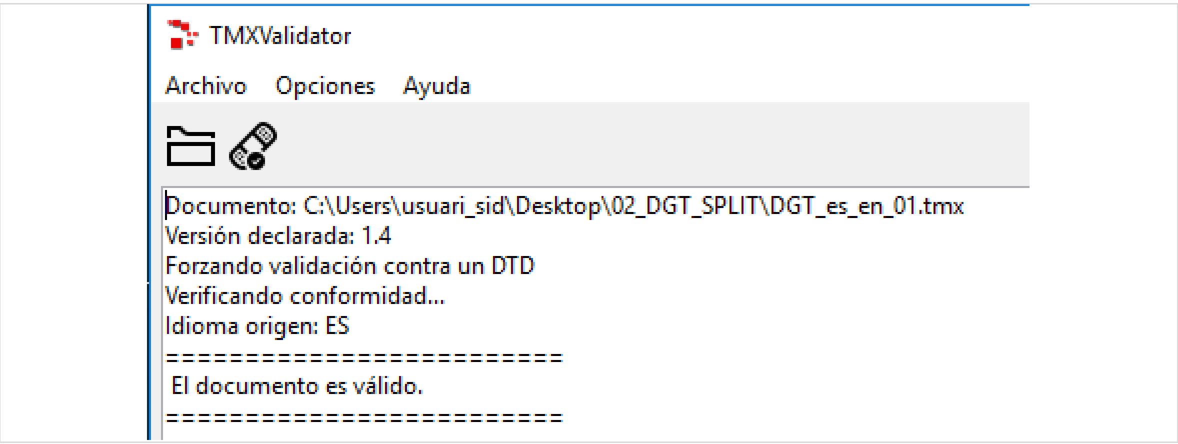
Fields: □ Key □ Flag □ ID □ Match □ Text_EN □ Text_ES

□ Match case
□ Match whole word
□ Use regular expressions

□ Reset flags before executing the selections

EL PRESIDENTE DE IRLANDA,

Appendix 4. TM confirms that the recently created TMs fulfill the .tmx standard.

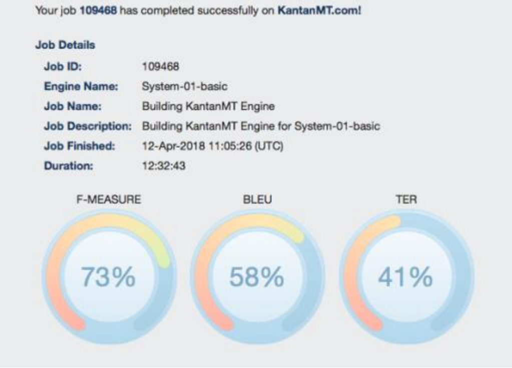


```
Documento: C:\Users\usuari_sid\Desktop\02_DGT_SPLIT\DGT_es_en_01.tmx
Versión declarada: 1.4
Forzando validación contra un DTD
Verificando conformidad...
Idioma origen: ES
==========================
 El documento es válido.
==========================
```

Appendix 5. Checklist of training data before building a PB-SMT in the KantanMT platform

Are your files less than 512MB?
Are your files saved in UTF-8 with Unix line breaks?
Are your TMX using the correct source and target language combination?
Did you open your recently aligned translation memories to check for any tags or codes that could pollute the training datasets?
Is your monolingual corpus saved as “source.utf8.trg.mono.”?
Does your glossary include the corresponding source and target language ISO codes on top of the glossary as shown in Figure 18?
Is your glossary saved as “glossary.xlsx”?

Appendix 6. KantanMT notifies when the training job is completed



Your job **109468** has completed successfully on **KantanMT.com!**

Job Details

Job ID: 109468
Engine Name: System-01-basic
Job Name: Building KantanMT Engine
Job Description: Building KantanMT Engine for System-01-basic
Job Finished: 12-Apr-2018 11:05:26 (UTC)
Duration: 12:32:43

Appendix 7. Rejected Words Report of System 1

File name	**Total Words**	**Rejected Words**	**Percentage**
DGT_es_en_06.tmx.tmx.noprop.tmx	2.093.516	198.212	9%
ECB_es_en.tmx.tmx.noprop.tmx	3.514.411	652.527	19%
News_es_en.tmx.tmx.noprop.tmx	6.286.578	96.093	2%
DGT_es_en_01.tmx.tmx.noprop.tmx	7.832.778	737.678	9%
DGT_es_en_05.tmx.tmx.noprop.tmx	8.839.176	704.228	8%
DGT_es_en_03.tmx.tmx.noprop.tmx	9.399.880	762.876	8%
DGT_es_en_04.tmx.tmx.noprop.tmx	9.836.761	809.472	8%
DGT_es_en_02.tmx.tmx.noprop.tmx	10.316.145	934.136	9%
		Average of Rejected Words	**9%**

Appendix 8. Source testing text. Excerpts taken from the Registro Nacional and MEIC

Texto 1: Las preguntas más frecuentes

3. ¿Cuáles son los principales beneficios que tendrán los nuevos empresarios con CrearEmpresa?

CrearEmpresa pretende apoyar a los interesados en la creación de su empresa y poner a su alcance los servicios del Estado, en forma fácil y expedita.

En una sola ventanilla electrónica, los interesados podrán realizar todos los trámites necesarios en línea para poder constituir y poner en operación su empresa o negocio.

Los trámites que se pueden realizar a través de CrearEmpresa son:

a. Inscripción de sociedades mercantiles en el Registro Nacional.
b. Legalización electrónica de libros sociales en el Registro Nacional.
c. Publicación de edictos en el diario oficial, La Gaceta.
d. Certificación de personería jurídica del Registro Nacional.
e. Certificación de propiedad de bienes inmuebles.
f. Certificación de planos catastrados digitalizados.
g. Certificado de uso de suelo.
h. Obtención del permiso sanitario de funcionamiento por parte del Ministerio de Salud.
i. Obtención de la Viabilidad Ambiental por parte de la Secretaría Técnica Nacional Ambiental (SETENA) para empresas de bajo impacto ambiental.
j. Certificado Veterinario de Operación del Servicio Nacional de Salud Animal (SENASA) del Ministerio de Agricultura y Ganadería.
k. Verificación de la póliza de riesgos de trabajo del Instituto Nacional de Seguros (INS).
l. Patente comercial.
m. Inscripción como patrono en la Caja Costarricense de Seguro Social.
n. Inscripción como contribuyente en la Dirección General de Tributación.

5. ¿Qué es la inscripción de una empresa?

La inscripción de una empresa es la creación jurídica de una nueva sociedad ante el Registro Nacional. Este trámite lo puede realizar únicamente el notario.

Texto 2: Procedimiento simplificado para extranjeros

REQUERIMIENTO B

(Fecha)

Señor
Director General
Dirección General de Migración y Extranjería de Costa Rica

Estimado Señor:

La compañía (nombre formal de la compañía), solicita autorización de estancia como visitante de negocios a favor del Señor(a) (nombre), mayor, (estado civil), Agente de Negocios, vecino de (dirección en país de origen), de paso por (dirección en Costa Rica), Costa Rica, ciudadano (nacionalidad), con pasaporte de su país número: (No. de pasaporte), que desea hacer negocios en Costa Rica durante un año en nombre de la compañía (nombre formal de la compañía), así como registrarse como proveedor y participar en licitaciones públicas a través del sistema de compras públicas Mer-link. Asimismo dicha persona no devengará el pago de salarios u honorarios en el país. Hace la solicitud de conformidad con el artículo 88, inciso 2) de la Ley General de Migración y Extranjería 8764, vigente a partir del 01 de marzo de 2010, cumpliendo con todos los requisitos.

Atentamente,

(Nombre Completo)
(Puesto: Presidente Ejecutivo, Representante Legal)
(Compañía)

NOMBRE APODERADO/DATOS ID

Acepto poder:

Texto 3: Registro Nacional de sociedades

A partir de la publicación, el interesado tiene 30 días hábiles para ejercer su oposición en Sede Judicial.

Toda información debe ser corroborada en La Gaceta y el Alcance correspondiente.

Método para consulta:

La forma más sencilla de ubicar una sociedad, es utilizando la tecla "BUSCAR" y digitando el número de consecutivo de la cédula jurídica.

No es necesario indicar los números correspondientes a tipo de entidad, a saber 3-101- 3-012- etc., sólo el consecutivo.

Cuando el consecutivo tenga menos de 6 dígitos, no debe digitarse ceros a la izquierda, únicamente el número que lo compone. Eje: 3-101-001234 (solo debe digitarse 1234).

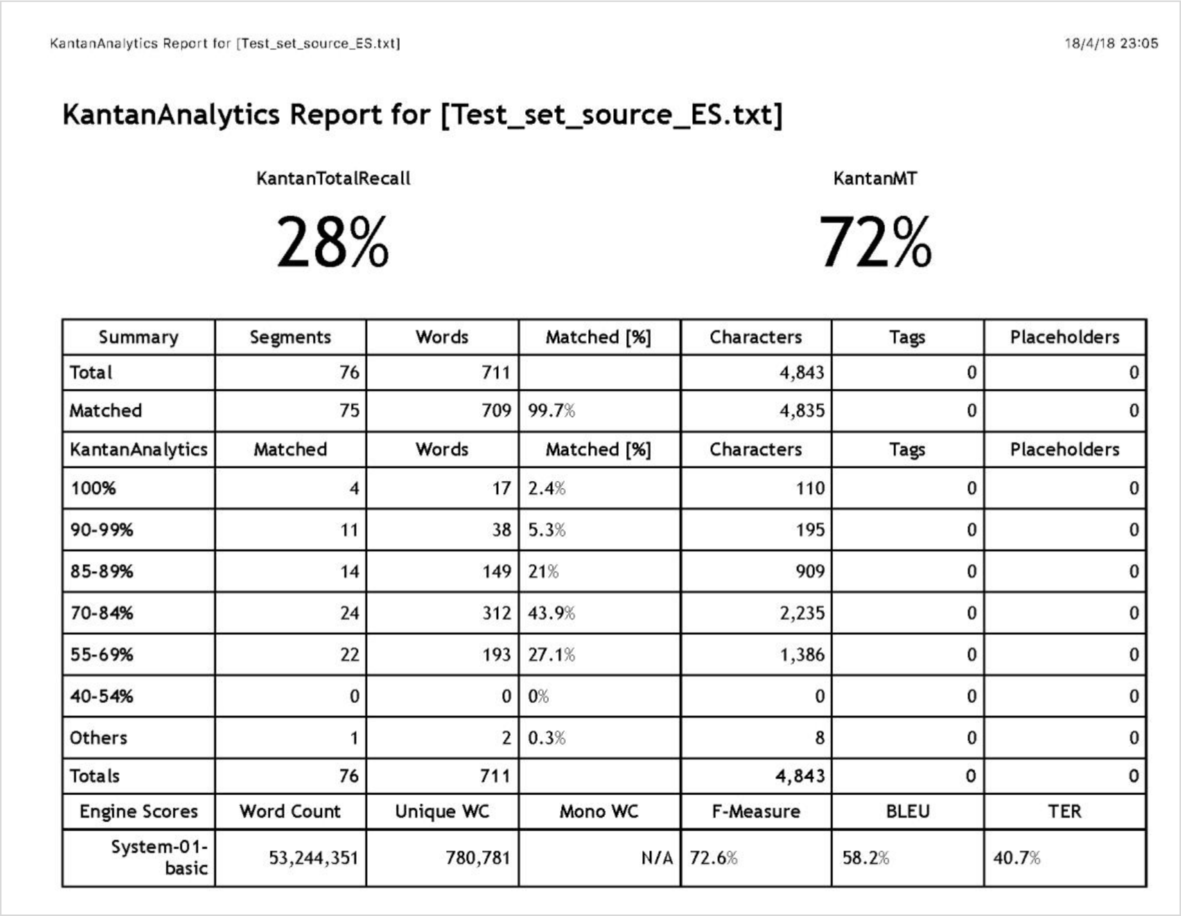
Texto 4: Servicios

En este registro se inscriben: empresas individuales de responsabilidad limitada, sociedades en nombre colectivo, sociedades en comandita, sociedades de responsabilidad limitada, sociedades anónimas; sociedades reguladas por leyes especiales, bolsa de valores, puestos de bolsa, sociedades de inversión, sociedades operadoras de fondos de inversión, centrales de valores, sociedades de compensación y liquidación, sociedades calificadoras de riesgo, sociedades operadoras de fondos de pensiones complementarias, empresas financieras de carácter no bancario, bancos privados, sociedades anónimas laborales, sociedades comercializadoras de seguros y sociedades anónimas deportivas.

Inscripción de fundaciones; sociedades de actividades profesionales; asociaciones civiles y deportivas. Inscripción de reformas a las personas jurídicas indicadas; también inscripción de poderes, otorgados por personas físicas; conferidos por sociedades nacionales; otorgados en el extranjero y conferidos por sociedades extranjeras; apertura de sucursales de empresas extranjeras.

Asimismo, se inscribe la insolvencia; insania; tutoría; albaceazgos; corredores jurados; reserva de nombre y además, fiscalización de asociaciones civiles; así como consulta de la información en el sistema automatizado y en el tradicional de tomos.

Appendix 9. KantanAnalytics Report for System 1



KantanAnalytics Report for [Test_set_source_ES.txt] 18/4/18 23:05

KantanAnalytics Report for [Test_set_source_ES.txt]

KantanTotalRecall

28%

KantanMT

72%

Summary	Segments	Words	Matched [%]	Characters	Tags	Placeholders
Total	76	711		4,843	0	0
Matched	75	709	99.7%	4,835	0	0
KantanAnalytics	Matched	Words	Matched [%]	Characters	Tags	Placeholders
100%	4	17	2.4%	110	0	0
90-99%	11	38	5.3%	195	0	0
85-89%	14	149	21%	909	0	0
70-84%	24	312	43.9%	2,235	0	0
55-69%	22	193	27.1%	1,386	0	0
40-54%	0	0	0%	0	0	0
Others	1	2	0.3%	8	0	0
Totals	76	711		4,843	0	0
Engine Scores	Word Count	Unique WC	Mono WC	F-Measure	BLEU	TER
System-01-basic	53,244,351	780,781	N/A	72.6%	58.2%	40.7%

Appendix 10. Unknown words that System 1 Fails to Recognize

1 **CrearEmpresa** ¿Cuáles son los principales beneficios que tendrán los nuevos empresarios con CrearEmpresa? What are the main benefits that will the new entrepreneurs with CrearEmpresa?

2 **personería** Certificación de personería jurídica del Registro Nacional. Certification of Legal personería the National Register.

3 **catastrados** Certificación de planos catastrados digitalizados. Certification of planes catastrados digitised.

4 **SETENA** Obtención de la Viabilidad Ambiental por parte de la Secretaría Técnica Nacional Ambiental (SETENA) para empresas de bajo impacto ambiental. Establishment of environmental sustainability by the Technical Secretariat national environmental (SETENA) for companies with a low environmental impact.

5 **de** **Certificado Veterinario de Operación del Servicio Nacional de Salud Animal (SENASA) del Ministerio de Agricultura y Ganadería**. Veterinary Certificate of Operation of the National Health Service (SENASA) Of the Animal, Ministerio de Agricultura y Ganadería.

6 **INS** **Verificación de la póliza de riesgos de trabajo del Instituto Nacional de Seguros (INS).** Verification of the insurance risk of the work of the National Institute for Insurance (INS).

7 **Social** **Inscripción como patrono en la Caja Costarricense de Seguro Social.** Registration As an Employer in the Box Squirrel of Social Insurance.

8 **General** **Inscripción como contribuyente en la Dirección General de Tributación.** Registration As a Taxpayer in the Directorate-General For Taxation.

9 **Director** Director General General Director

10 **formal** **La compañía (nombre formal de la compañía**), solicita autorización de estancia como visitante de negocios a favor del Señor(a) (nombre), mayor, (estado civil), Agente de Negocios, vecino de (dirección en país de origen), de paso por (dirección en Costa Rica), Costa Rica, ciudadano (nacionalidad), con pasaporte de su país número: The company (formal name of the carrier), request authorisation for stay as visiting business Mr (a) in favour of (name), higher (State civil), staff turnover, neighbor (address in the country of origin), crossing (address in Costa Rica), Costa Rica, citizen (nationality), with their country's passport number:

11 **No** **(No. de pasaporte),** que desea hacer negocios en Costa Rica durante un año en nombre de la compañía (nombre formal de la compañía), así como registrarse como proveedor y participar en licitaciones públicas a través del sistema de compras públicas Mer-link. (No passport), which wishes to do business in Costa Rica for a year on behalf of the company (the formal name of the carrier), as well as be recorded as the supplier and participate in public tenders through the system of public purchasing foreign-link.

12 **Legal** Presidente Ejecutivo, Representante Legal) Executive President, Legal Representative)

13 **ID** **NOMBRE APODERADO/DATOS ID** NAME PROXY/INFORMATION ID

14 **Judicial** A partir de la publicación, el interesado tiene 30 días hábiles para ejercer su oposición en Sede Judicial. From the date of publication, the person concerned has 30 working days in order to exercise its opposition to Judicial office.

15 **digitando** La forma más sencilla de ubicar una sociedad, es utilizando la tecla "BUSCAR" y digitando el número de consecutivo de la cédula jurídica. The simplest way to locate a society, it is using the key "seek" and digitando the number of consecutive instituting the legal entity.

16 **etc** No es necesario indicar los números correspondientes a tipo de entidad, a saber 3-101- 3-012- etc., sólo el consecutivo. It is not necessary to indicate the numbers corresponding to the type of entity, namely 3-101- 3-012- etc., only the row.

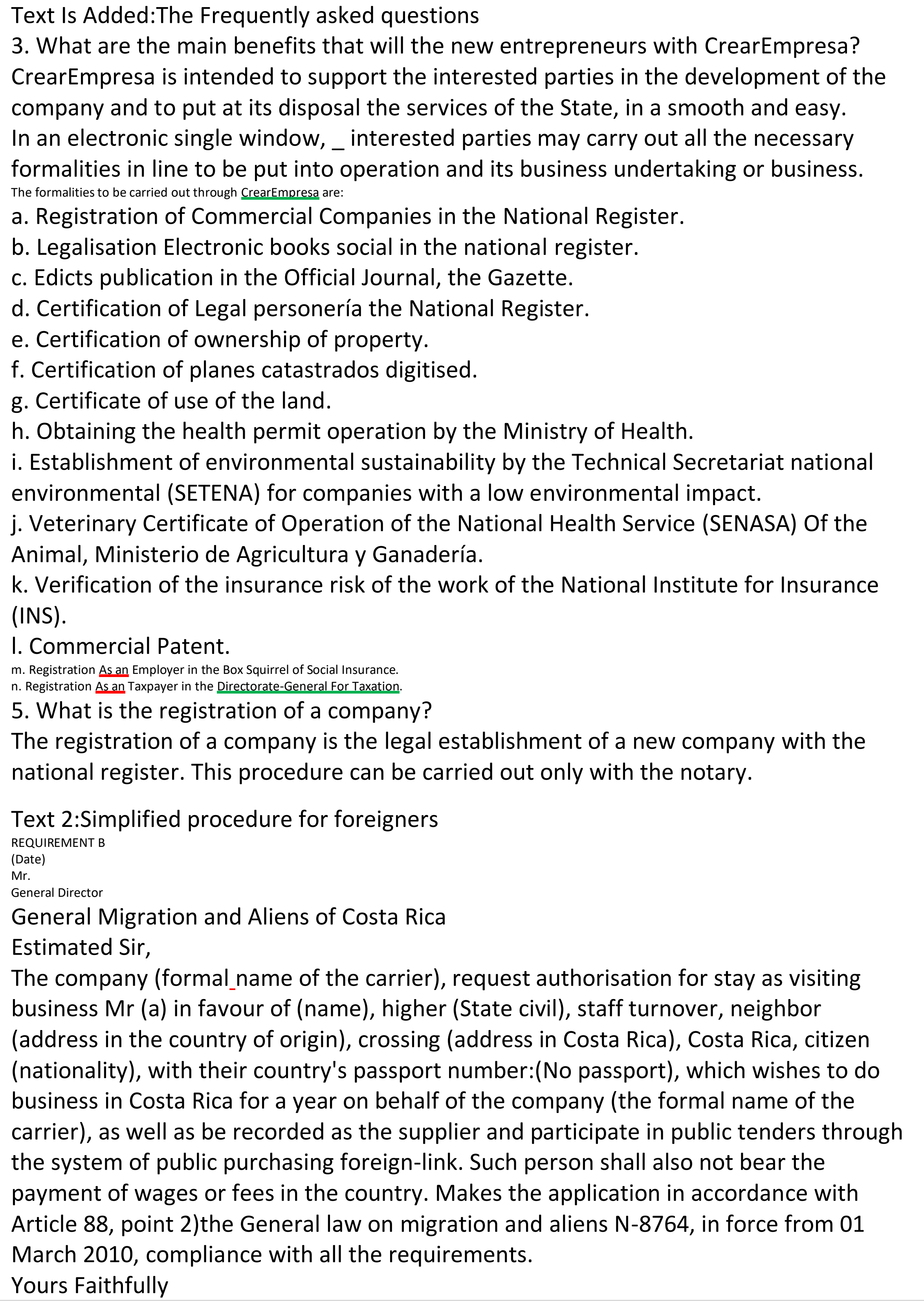
17 **digitarse** Cuando el consecutivo tenga menos de 6 dígitos, no debe digitarse ceros a la izquierda, únicamente el número que lo compone. When the sequential has six digits, should not digitarse zeros to the left, only the number it is composed.

18 **inscriben** En este registro se inscriben: In this record inscriben:

19 **operadoras** empresas individuales de responsabilidad limitada, sociedades en nombre colectivo, sociedades en comandita, sociedades de responsabilidad limitada, sociedades anónimas; sociedades reguladas por leyes especiales, bolsa de valores, puestos de bolsa, sociedades de inversión, sociedades operadoras de fondos de inversión, centrales de valores, sociedades de compensación y liquidación, sociedades calificadoras de riesgo, sociedades operadoras de fondos de pensiones complementarias, empresas financieras de carácter no bancario, bancos privados, sociedades anónimas laborales, sociedades comercializadoras de seguros y sociedades anónimas deportivas. Individual companies with limited liability companies, in collective name, companies limited partnership, Limited liability company, corporation donations; companies covered by special laws, stock exchange, stock exchange positions, investment trusts, companies operadoras of investment funds, central securities clearing and settlement companies, qualifying companies, the risk of corporate operadoras supplementary pension funds, finance companies non-banking, private banks, corporations work, traders and insurance companies limited companies and sporting activities.

20 **indicadas** Inscripción de reformas a las personas jurídicas indicadas; también inscripción de poderes, otorgados por personas físicas; conferidos por sociedades nacionales; otorgados en el extranjero y conferidos por sociedades extranjeras; apertura de sucursales de empresas extranjeras. Registration of reforms to legal persons indicadas; also registration of powers, awarded by natural persons; conferred by national companies granted abroad and conferred by foreign companies'stocks; companies open a branch of foreign companies.

Appendix 11. Output text of System 1



Text Is Added:The Frequently asked questions

3. What are the main benefits that will the new entrepreneurs with CrearEmpresa?

CrearEmpresa is intended to support the interested parties in the development of the company and to put at its disposal the services of the State, in a smooth and easy.

In an electronic single window, _ interested parties may carry out all the necessary formalities in line to be put into operation and its business undertaking or business.

The formalities to be carried out through CrearEmpresa are:

a. Registration of Commercial Companies in the National Register.

b. Legalisation Electronic books social in the national register.

c. Edicts publication in the Official Journal, the Gazette.

d. Certification of Legal personería the National Register.

e. Certification of ownership of property.

f. Certification of planes catastrados digitised.

g. Certificate of use of the land.

h. Obtaining the health permit operation by the Ministry of Health.

i. Establishment of environmental sustainability by the Technical Secretariat national environmental (SETENA) for companies with a low environmental impact.

j. Veterinary Certificate of Operation of the National Health Service (SENASA) Of the Animal, Ministerio de Agricultura y Ganadería.

k. Verification of the insurance risk of the work of the National Institute for Insurance (INS).

l. Commercial Patent.

m. Registration As an Employer in the Box Squirrel of Social Insurance.

n. Registration As an Taxpayer in the Directorate-General For Taxation.

5. What is the registration of a company?

The registration of a company is the legal establishment of a new company with the national register. This procedure can be carried out only with the notary.

Text 2:Simplified procedure for foreigners

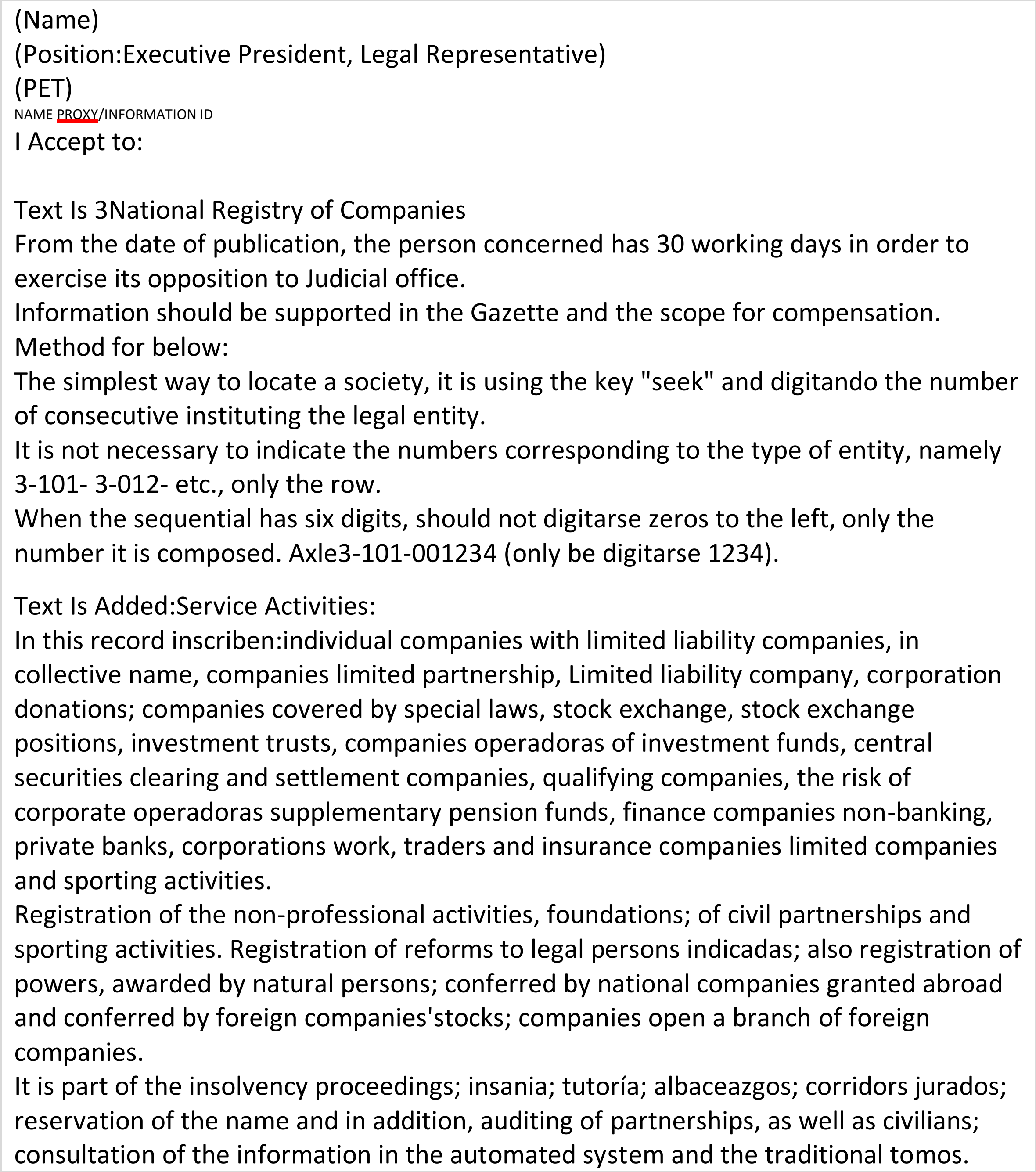
REQUIREMENT B
(Date)
Mr.
General Director

General Migration and Aliens of Costa Rica

Estimated Sir,

The company (formal name of the carrier), request authorisation for stay as visiting business Mr (a) in favour of (name), higher (State civil), staff turnover, neighbor (address in the country of origin), crossing (address in Costa Rica), Costa Rica, citizen (nationality), with their country's passport number:(No passport), which wishes to do business in Costa Rica for a year on behalf of the company (the formal name of the carrier), as well as be recorded as the supplier and participate in public tenders through the system of public purchasing foreign-link. Such person shall also not bear the payment of wages or fees in the country. Makes the application in accordance with Article 88, point 2)the General law on migration and aliens N-8764, in force from 01 March 2010, compliance with all the requirements.

Yours Faithfully



(Name)
(Position:Executive President, Legal Representative)
(PET)
NAME PROXY/INFORMATION ID
I Accept to:

Text Is 3National Registry of Companies
From the date of publication, the person concerned has 30 working days in order to exercise its opposition to Judicial office.
Information should be supported in the Gazette and the scope for compensation.
Method for below:
The simplest way to locate a society, it is using the key "seek" and digitando the number of consecutive instituting the legal entity.
It is not necessary to indicate the numbers corresponding to the type of entity, namely 3-101- 3-012- etc., only the row.
When the sequential has six digits, should not digitarse zeros to the left, only the number it is composed. Axle3-101-001234 (only be digitarse 1234).

Text Is Added:Service Activities:
In this record inscriben:individual companies with limited liability companies, in collective name, companies limited partnership, Limited liability company, corporation donations; companies covered by special laws, stock exchange, stock exchange positions, investment trusts, companies operadoras of investment funds, central securities clearing and settlement companies, qualifying companies, the risk of corporate operadoras supplementary pension funds, finance companies non-banking, private banks, corporations work, traders and insurance companies limited companies and sporting activities.
Registration of the non-professional activities, foundations; of civil partnerships and sporting activities. Registration of reforms to legal persons indicadas; also registration of powers, awarded by natural persons; conferred by national companies granted abroad and conferred by foreign companies'stocks; companies open a branch of foreign companies.
It is part of the insolvency proceedings; insania; tutoría; albaceazgos; corridors jurados; reservation of the name and in addition, auditing of partnerships, as well as civilians; consultation of the information in the automated system and the traditional tomos.

Appendix 12. Example List of Entries Added to the Run-time Glossary

domicilio	domicile
notario	notary
Dirección General de Tributación	Directorate-General for Taxation
patente comercial	patents
uso de suelo	land use
planos catastrados digitalizados	digitized cadastral plans

inscripción	registration
suscribirse	to register
albaceazgos	Executorships
legalización	legalization
certificado veterinario de operación	certificate for veterinary operations
poliza de riesgos de trabajo	insurance policy for occupational risk
authorisation	authorization
permiso sanitario	sanitary permit
viabilidad ambiental	environmental feasibility
agente de negocios	business agent
Sede Judicial	court
Texto	Text

Appendix 13. Example List of Entries Added to the Training Glossary

en línea	online
en línea con	in line with
publicación de	publication of
obtención	issuance
empresa	business
compañía	company
verificación	verification
sociedad	company
registro nacional de	national registry of
puerto	port
puerto	Puerto
estimado	dear
estimado	estimate
formal	official

Appendix 14. Account on the Attempt to Train a NMT engine, in the AWS Deep Learning AMIs

To run a NMT engine, bear in mind the following: 1) the translator must have an intermediate knowledge on how to use the command line to train the engine, 2) the support of a software engineer is still required to install and prepare the AMIs, 3) the infrastructure to run the MMT software is very expensive.

We followed Andy Way's motto (engineers and translators should come together), so an engineer participated in the launch and preparation of an AWS instance, which is basically a cloud-based server. Another characteristic of this technology is that the user can select the kind of environment. In this case, we used the AWS Deep Learning AMIs[56], a dedicated instance to run deep learning technologies. One of the advantages of this specific AMIs is that it comes with a set of pre-configured frameworks to run the MMT software such as TensorFlow and NVIDIA's GPU drivers. Once the instance had been launched, the engineer proceeded with the installation of the rest of packages necessary to run the MMT software, like Java 8 libraries, PyTorch 0.3 and MMT Binaries[57].

SETTING UP THE INSTANCE TO RUN THE MMT ENGINE

To begin with the training process, it is first necessary to upload the training datasets to the instance; for this, the engineer used the following file transfer protocol:

File transfer protocol = SFTP-3

Cryptographic protocol = SSH-2

SSH implementation = OpenSSH_7.2p2 Ubuntu-4ubuntu2.4

Encryption algorithm = aes

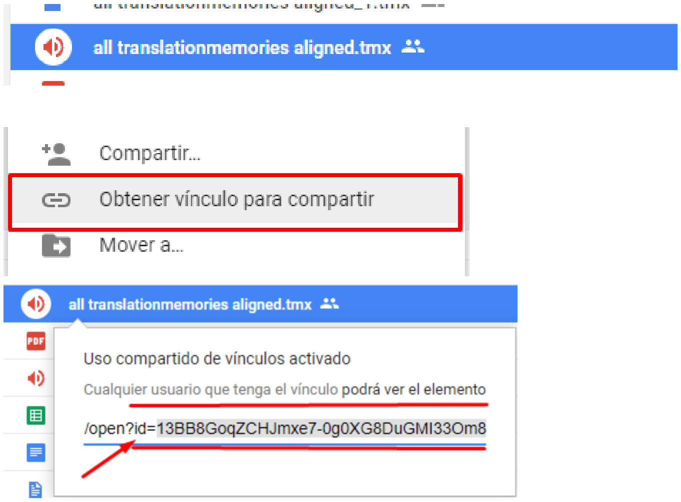
Compression = No

Even though the instance had enough storage capacity and network throughout to download large files, the SFTP protocol failed due to the lack of fault-tolerance of the protocol itself, so the files got corrupted and inconsistent, thus unusable for training purposes. As a workaround, the files were uploaded to Google Drive, and the

[56] https://aws.amazon.com/machine-learning/amis/?nc1=h_ls

[57] Read more about the MMT installation requirements and processes https://github.com/ModernMT/MMT/blob/master/INSTALL.md

corresponding sharable link was entered within a Python script that could manage large-files downloads. The script was then executed within the AWS instance.


all translationmemories aligned.tmx
Compartir...
Obtener vínculo para compartir
Mover a...
all translationmemories aligned.tmx
PDF

Uso compartido de vínculos activado

Cualquier usuario que tenga el vínculo podrá ver el elemento

/open?id=13BB8GoqZCHJmxe7-0g0XG8DuGMI33Om8

The ID number shown above is used in the following command:

```
(python2) Ubuntu@up-172-31-66-218:~ Python download_gdrive.py
13BB8GoqZCHJmxe7-0g0XG8DuGMI33Om8
downloads/all_translationmemories_aligned.tmx
```

This command helps to download the dataset into the instance, and it indicates: 1) the name of the script and 2) the ID number of the datasets previously uploaded into Google Drive.

THE TRAINING WITH THE SAMPLE FILES

Once these preparation and launching steps were performed, it was time to begin the training process. We first opened the command line locally (from our computer' terminal), and run the following commands to begin the session within the Ubuntu instance:

```
cd ~/Downloads
chmod 400 gloriana_kp.pem
ssh -i "gloriana_kp.pem" ubuntu@18.232.220.36
```

Using those commands, we accessed the Ubuntu instance under the name of (python2)Ubuntu@ip-172-31-66-218.

Second, we ran the following command to create a new engine named ***en-it-Normal***:

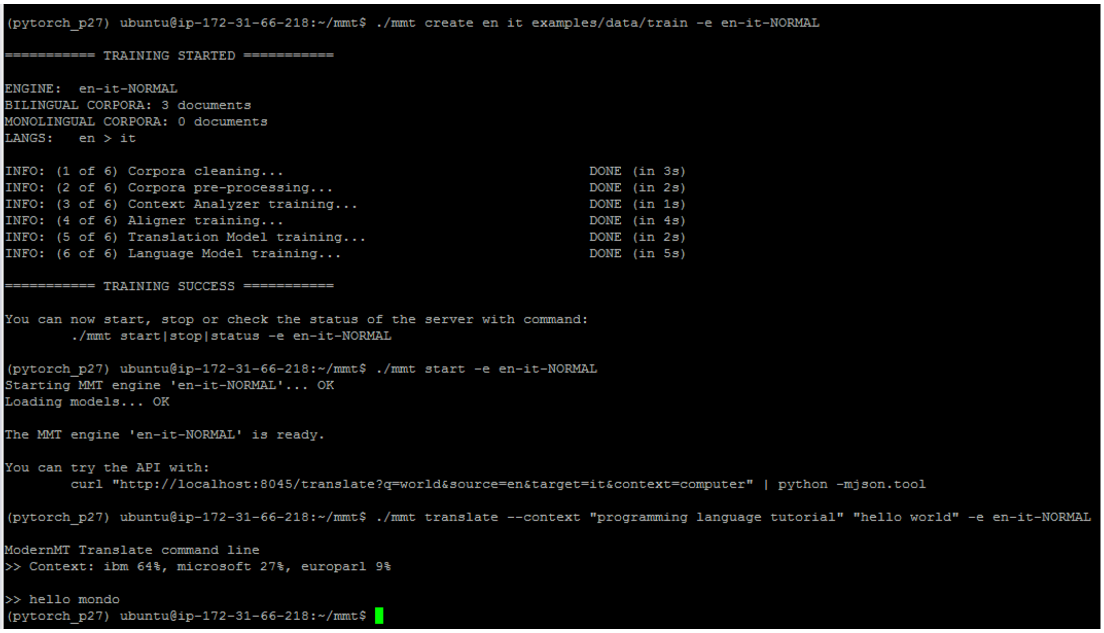
```
$ ./mmt create en it examples/data/train –e en-it-NORMAL
```

Then, to run the training using the examples dataset, we ran the following command:

```
$ ./mmt start
```

This engine was trained on the sample files found within the MMT's github account. They are in total: 1,6 MB.

Example Training Dataset Downloaded from the MMT's Github page.	
Europarl.en Europarl.it	304 KB 335 KB
Ibm.en Ibm.it	153 KB 176 KB
Microsoft.en Microsoft.it	290 KB 341 KB

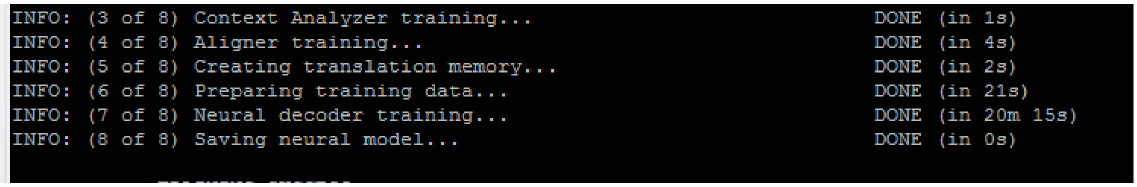
t 

```
(pytorch_p27) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt create en it examples/data/train -e en-it-NORMAL

=========== TRAINING STARTED ===========

ENGINE:  en-it-NORMAL
BILINGUAL CORPORA: 3 documents
MONOLINGUAL CORPORA: 0 documents
LANGS:   en > it

INFO: (1 of 6) Corpora cleaning...                                      DONE (in 3s)
INFO: (2 of 6) Corpora pre-processing...                                DONE (in 2s)
INFO: (3 of 6) Context Analyzer training...                             DONE (in 1s)
INFO: (4 of 6) Aligner training...                                      DONE (in 4s)
INFO: (5 of 6) Translation Model training...                            DONE (in 2s)
INFO: (6 of 6) Language Model training...                               DONE (in 5s)

=========== TRAINING SUCCESS ===========

You can now start, stop or check the status of the server with command:
        ./mmt start|stop|status -e en-it-NORMAL

(pytorch_p27) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt start -e en-it-NORMAL
Starting MMT engine 'en-it-NORMAL'... OK
Loading models... OK

The MMT engine 'en-it-NORMAL' is ready.

You can try the API with:
        curl "http://localhost:8045/translate?q=world&source=en&target=it&context=computer" | python -mjson.tool

(pytorch_p27) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt translate --context "programming language tutorial" "hello world" -e en-it-NORMAL

ModernMT Translate command line
>> Context: ibm 64%, microsoft 27%, europarl 9%

>> hello mondo
(pytorch_p27) ubuntu@ip-172-31-66-218:~/mmt$ 
```

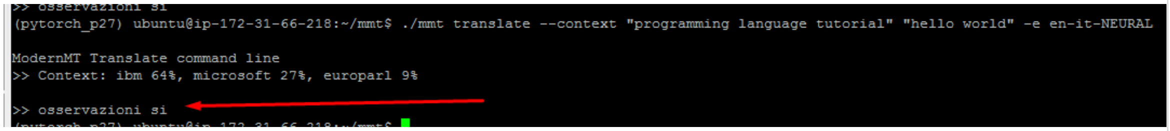
The training took approximately 17 seconds. Finally, we were able to test our engine by running the following command:

```
$ ./mmt translate --context "programming langauge tutorial" "hello world"—e en-it-Normal
```

To this instruction the engine successfully translates "hello world" for "hello mondo." After having experimented with the PB-SMT training process, we followed the corresponding steps to train a Neural MT. We found out that under the same circumstances the training takes 00:20:17:00, and that the translation was not correct. It translates “hello mondo” for “osservazioni si”. See the screenshots below:



```
INFO: (3 of 8) Context Analyzer training...                        DONE (in 1s)
INFO: (4 of 8) Aligner training...                                 DONE (in 4s)
INFO: (5 of 8) Creating translation memory...                      DONE (in 2s)
INFO: (6 of 8) Preparing training data...                          DONE (in 21s)
INFO: (7 of 8) Neural decoder training...                          DONE (in 20m 15s)
INFO: (8 of 8) Saving neural model...                              DONE (in 0s)
```



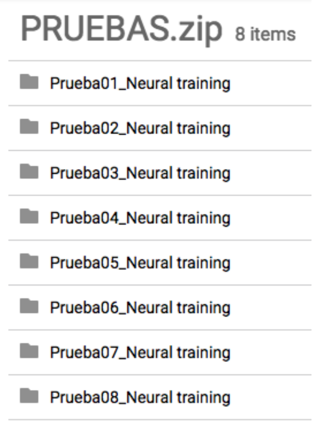
```
>> osservazioni si
(pytorch_p27) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt translate --context "programming language tutorial" "hello world" -e en-it-NEURAL

ModernMT Translate command line
>> Context: ibm 64%, microsoft 27%, europarl 9%

>> osservazioni si
(pytorch_p27) ubuntu@ip-172-31-66-218:~/mmt$
```

We carried out the corresponding “on the fly” corrections that MMT engines feature and were able to obtain the correct translation of “hello mondo”. At that point, we resolved that the reason why the neural training was taking so long was due to the fact that we had ran it within a CPU instance; thus, we would have needed to create and launch a GPU instance to run a neural training.

To this, a new problem arose. GPU instances are not that affordable. A CPU instance costs $0.1 per hour, while a GPU one costs $1 per hour, meaning 10 times more expensive. Taking into consideration that the training of 1,6 MB takes 20’, a training of 160 MB could take 2000’ meaning 33 hours. The cost factor was definitely a discouraging reason why training a neural MT was not pursuable any more for an average translator. Alternative options to this was the KantanMT, but for the same reasons we couldn’t continue with our primary objective. The monthly subscription to Kantan Neural features is of 999 euros.

Appendix 15. An Account on Training an NMT Engine Using the Custom Datasets of this Use-Case

One of the outcomes from this experimental phase was to find out that MMT struggles much more to train on TMX files than with corpora files. In a testing phase, 8 zipped folders containing 2 sets of folders with a TMX file and the corresponding corpora were uploaded to the AWS instance.
All trainings containing a TMX files failed, whereas the tests with corpora, even big files were successfully uploaded. See the following screenshot and their respective result.



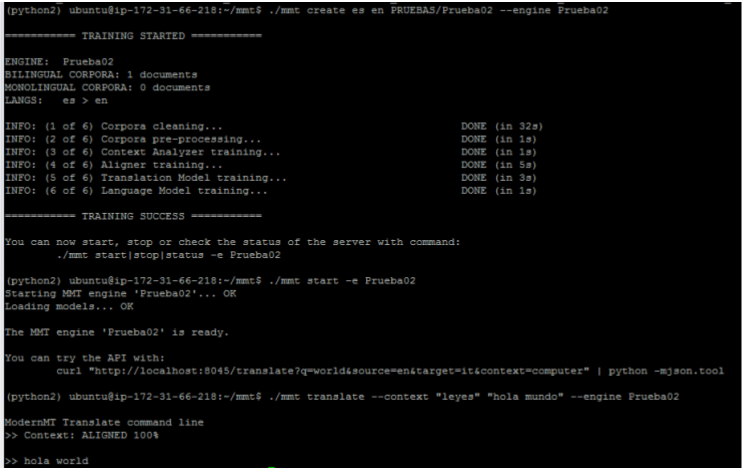
PRUEBAS.zip 8 items

- Prueba01_Neural training
- Prueba02_Neural training
- Prueba03_Neural training
- Prueba04_Neural training
- Prueba05_Neural training
- Prueba06_Neural training
- Prueba07_Neural training
- Prueba08_Neural training

FINAL RESULT: All trainings containing TMX files failed, whereas the tests with corpora, even big files were successfully uploaded. See the following screenshot and their respective result.

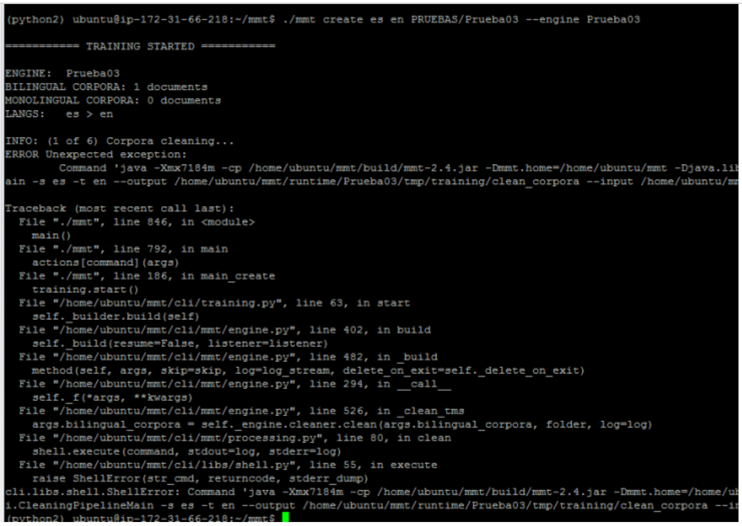
TEST 1: TMX FILE
R:/The training failed in step 1.

TEST 2: CORPUS
R:/Training Success



```
(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt create es en PRUEBAS/Prueba02 --engine Prueba02

=========== TRAINING STARTED ===========

ENGINE:  Prueba02
BILINGUAL CORPORA: 1 documents
MONOLINGUAL CORPORA: 0 documents
LANGS:   es > en

INFO: (1 of 6) Corpora cleaning...                                     DONE (in 32s)
INFO: (2 of 6) Corpora pre-processing...                               DONE (in 1s)
INFO: (3 of 6) Context Analyzer training...                            DONE (in 1s)
INFO: (4 of 6) Aligner training...                                     DONE (in 5s)
INFO: (5 of 6) Translation Model training...                           DONE (in 3s)
INFO: (6 of 6) Language Model training...                              DONE (in 1s)

=========== TRAINING SUCCESS ===========

You can now start, stop or check the status of the server with command:
        ./mmt start|stop|status -e Prueba02

(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt start -e Prueba02
Starting MMT engine 'Prueba02'... OK
Loading models... OK

The MMT engine 'Prueba02' is ready.

You can try the API with:
        curl "http://localhost:8045/translate?q=world&source=en&target=it&context=computer" | python -mjson.tool

(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt translate --context "leyes" "hola mundo" --engine Prueba02

ModernMT Translate command line
>> Context: ALIGNED 100%

>> hola world
```

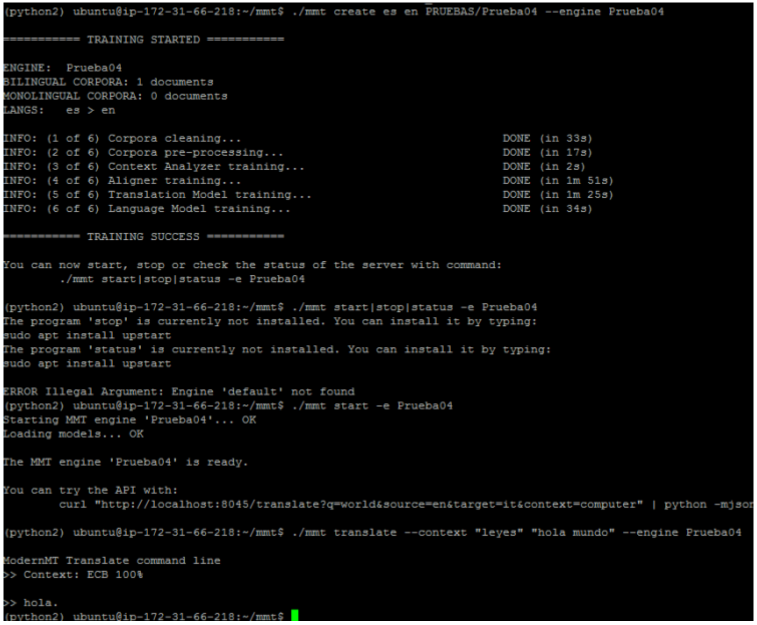
TEST 3: TMX FILE
R:/It failed.



```
(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt create es en PRUEBAS/Prueba03 --engine Prueba03

=========== TRAINING STARTED ===========

ENGINE:  Prueba03
BILINGUAL CORPORA: 1 documents
MONOLINGUAL CORPORA: 0 documents
LANGS:   es > en

INFO: (1 of 6) Corpora cleaning...
ERROR Unexpected exception:
        Command 'java -Xmx7184m -cp /home/ubuntu/mmt/build/mmt-2.4.jar -Dmmt.home=/home/ubuntu/mmt -Djava.lib
ain -s es -t en --output /home/ubuntu/mmt/runtime/Prueba03/tmp/training/clean_corpora --input /home/ubuntu/mm

Traceback (most recent call last):
  File "./mmt", line 846, in <module>
    main()
  File "./mmt", line 792, in main
    actions[command](args)
  File "./mmt", line 186, in main_create
    training.start()
  File "/home/ubuntu/mmt/cli/training.py", line 63, in start
    self._builder.build(self)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 402, in build
    self._build(resume=False, listener=listener)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 482, in _build
    method(self, args, skip=skip, log=log_stream, delete_on_exit=self._delete_on_exit)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 294, in __call__
    self._f(*args, **kwargs)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 526, in _clean_tms
    args.bilingual_corpora = self._engine.cleaner.clean(args.bilingual_corpora, folder, log=log)
  File "/home/ubuntu/mmt/cli/mmt/processing.py", line 80, in clean
    shell.execute(command, stdout=log, stderr=log)
  File "/home/ubuntu/mmt/cli/libs/shell.py", line 55, in execute
    raise ShellError(str_cmd, returncode, stderr_dump)
cli.libs.shell.ShellError: Command 'java -Xmx7184m -cp /home/ubuntu/mmt/build/mmt-2.4.jar -Dmmt.home=/home/ub
i.CleaningPipelineMain -s es -t en --output /home/ubuntu/mmt/runtime/Prueba03/tmp/training/clean_corpora --in
(python2) ubuntu@ip-172-31-66-218:~/mmt$ 
```

TEST 4: CORPUS

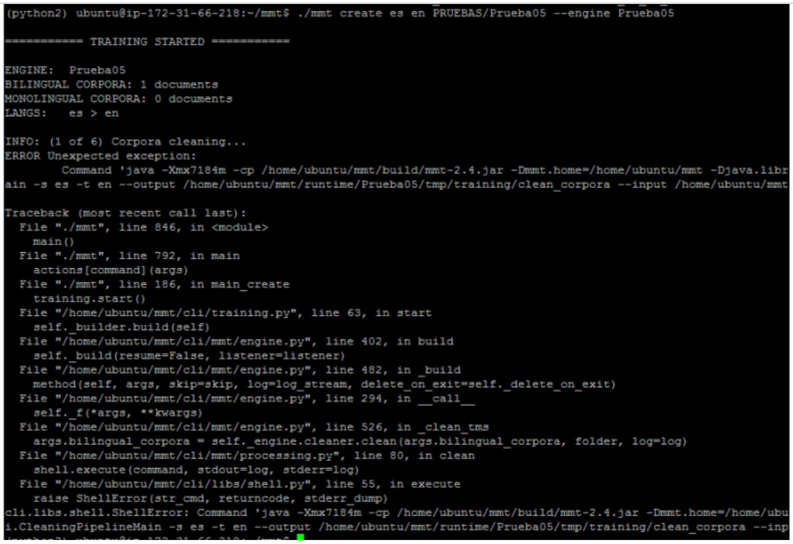
R:/ Training Success



```
(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt create es en PRUEBAS/Prueba04 --engine Prueba04

=========== TRAINING STARTED ===========

ENGINE:  Prueba04
BILINGUAL CORPORA: 1 documents
MONOLINGUAL CORPORA: 0 documents
LANGS:   es > en

INFO: (1 of 6) Corpora cleaning...                                     DONE (in 33s)
INFO: (2 of 6) Corpora pre-processing...                               DONE (in 17s)
INFO: (3 of 6) Context Analyzer training...                            DONE (in 2s)
INFO: (4 of 6) Aligner training...                                     DONE (in 1m 51s)
INFO: (5 of 6) Translation Model training...                           DONE (in 1m 25s)
INFO: (6 of 6) Language Model training...                              DONE (in 34s)

=========== TRAINING SUCCESS ===========

You can now start, stop or check the status of the server with command:
        ./mmt start|stop|status -e Prueba04

(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt start|stop|status -e Prueba04
The program 'stop' is currently not installed. You can install it by typing:
sudo apt install upstart
The program 'status' is currently not installed. You can install it by typing:
sudo apt install upstart

ERROR Illegal Argument: Engine 'default' not found
(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt start -e Prueba04
Starting MMT engine 'Prueba04'... OK
Loading models... OK

The MMT engine 'Prueba04' is ready.

You can try the API with:
        curl "http://localhost:8045/translate?q=world&source=en&target=it&context=computer" | python -mjson

(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt translate --context "leyes" "hola mundo" --engine Prueba04

ModernMT Translate command line
>> Context: ECB 100%

>> hola.
(python2) ubuntu@ip-172-31-66-218:~/mmt$ 
```

TEST 5: TMX

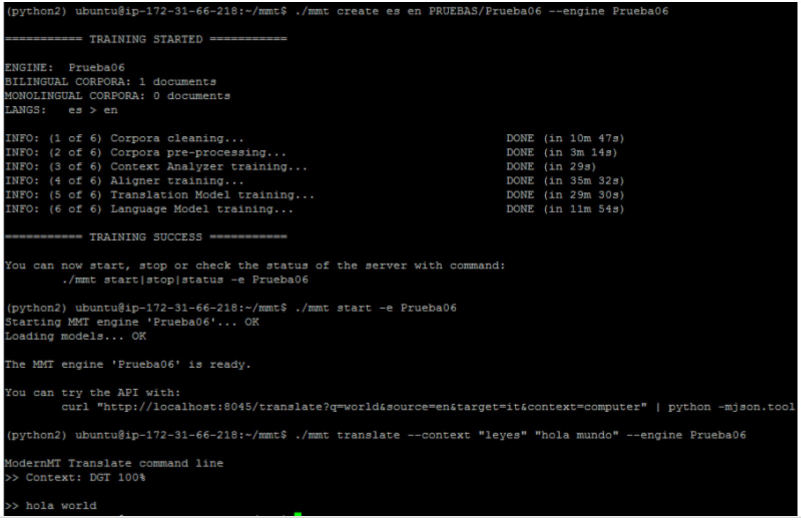
R:/ Failed.



```
(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt create es en PRUEBAS/Prueba05 --engine Prueba05

=========== TRAINING STARTED ===========

ENGINE:  Prueba05
BILINGUAL CORPORA: 1 documents
MONOLINGUAL CORPORA: 0 documents
LANGS:   es > en

INFO: (1 of 6) Corpora cleaning...
ERROR Unexpected exception:
        Command 'java -Xmx7184m -cp /home/ubuntu/mmt/build/mmt-2.4.jar -Dmmt.home=/home/ubuntu/mmt -Djava.libr
ain -s es -t en --output /home/ubuntu/mmt/runtime/Prueba05/tmp/training/clean_corpora --input /home/ubuntu/mmt

Traceback (most recent call last):
  File "./mmt", line 846, in <module>
    main()
  File "./mmt", line 792, in main
    actions[command](args)
  File "./mmt", line 186, in main_create
    training.start()
  File "/home/ubuntu/mmt/cli/training.py", line 63, in start
    self._builder.build(self)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 402, in build
    self._build(resume=False, listener=listener)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 482, in _build
    method(self, args, skip=skip, log=log_stream, delete_on_exit=self._delete_on_exit)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 294, in __call__
    self._f(*args, **kwargs)
  File "/home/ubuntu/mmt/cli/mmt/engine.py", line 526, in _clean_tms
    args.bilingual_corpora = self._engine.cleaner.clean(args.bilingual_corpora, folder, log=log)
  File "/home/ubuntu/mmt/cli/mmt/processing.py", line 80, in clean
    shell.execute(command, stdout=log, stderr=log)
  File "/home/ubuntu/mmt/cli/libs/shell.py", line 55, in execute
    raise ShellError(str_cmd, returncode, stderr_dump)
cli.libs.shell.ShellError: Command 'java -Xmx7184m -cp /home/ubuntu/mmt/build/mmt-2.4.jar -Dmmt.home=/home/ubu
i.CleaningPipelineMain -s es -t en --output /home/ubuntu/mmt/runtime/Prueba05/tmp/training/clean_corpora --inp
```

TEST 6: CORPUS OF APPROX. 700 MB

R:/ Training Success



```
(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt create es en PRUEBAS/Prueba06 --engine Prueba06

=========== TRAINING STARTED ===========

ENGINE:  Prueba06
BILINGUAL CORPORA: 1 documents
MONOLINGUAL CORPORA: 0 documents
LANGS:   es > en

INFO: (1 of 6) Corpora cleaning...                                      DONE (in 10m 47s)
INFO: (2 of 6) Corpora pre-processing...                                DONE (in 3m 14s)
INFO: (3 of 6) Context Analyzer training...                             DONE (in 29s)
INFO: (4 of 6) Aligner training...                                      DONE (in 35m 32s)
INFO: (5 of 6) Translation Model training...                            DONE (in 29m 30s)
INFO: (6 of 6) Language Model training...                               DONE (in 11m 54s)

=========== TRAINING SUCCESS ===========

You can now start, stop or check the status of the server with command:
        ./mmt start|stop|status -e Prueba06

(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt start -e Prueba06
Starting MMT engine 'Prueba06'... OK
Loading models... OK

The MMT engine 'Prueba06' is ready.

You can try the API with:
        curl "http://localhost:8045/translate?q=world&source=en&target=it&context=computer" | python -mjson.tool

(python2) ubuntu@ip-172-31-66-218:~/mmt$ ./mmt translate --context "leyes" "hola mundo" --engine Prueba06

ModernMT Translate command line
>> Context: DGT 100%

>> hola world
```

Appendix 16. Output Text of Google Translate

Text 2: Frequently asked questions

3. What are the main benefits that new entrepreneurs will have with CrearEmpresa?

CrearEmpresa aims to support those interested in the creation of their company and put the State services at their fingertips, easily and expeditiously.

In a single electronic window, the interested parties will be able to carry out all the necessary procedures to be able to set up online and put into operation their company or business.

The procedures that can be done through CrearEmpresa are:

to. Registration of mercantile companies in the National Registry.
b. Electronic legalization of social books in the National Registry.
c. Publication of edicts in the official newspaper La Gaceta.
d. Certification of legal status of the National Registry.
and. Certification of ownership of real estate.
F. Certification of digitized surveyed plans.
g. Land Use Certificate.
h. Obtaining the sanitary permission of operation on the part of the Ministry of Health.
i. Obtaining Environmental Viability by the National Environmental Technical Secretariat (SETENA) for companies with low environmental impact.
j. Veterinary Certificate of Operation of the National Animal Health Service (SENASA) of the Ministry of Agriculture and Livestock.
k. Verification of the work risks policy of the National Insurance Institute (INS).
l. Commercial patent.
m. Registration as employer in the Costa Rican Social Security Fund.
n. Registration as a taxpayer in the General Directorate of Taxation.

5. What is the registration of a company?

The registration of a company is the legal creation of a new company before the National Registry. This procedure can be done only by the Notary.

Text 3: Simplified procedure for foreigners

REQUIREMENT C
POWER

Ample and sufficient special power as in right corresponds so that on behalf of the Company COMPLETE NAME OF THE COMPANY.

The NAME OF THE AUTHORIZED PERSON can make the registration of this in the Electronic System of Recognition and in the Electronic Registry of Mer-link providers.

DIGITAL GOVERNMENT OF COSTA RICA:

Who we subscribe, Name, Sex, Nationality, adult, holder of passport number # Passport, in my capacity as Proxies of Company Name, limited company PAIS DE LA EMPRESA duly registered DATA OF THE REGISTRY OF THE COMPANY; Data of the Public Entity in which the REGISTRY is made, with address in the address of the company, duly authorized for this act as stated in DOCUMENT THAT ACCREDITED, by this means we confer Special Power in the Name of the Authorized, Sex, of nationality XXXXX, CIVIL STATUS, of legal age, with IDENTIFICATION NUMBER, this special power is so broad and sufficient as in the corresponding law so that on behalf of the company represented, NAME OF THE COMPANY.

can make the registration of this in the Electronic System of Recognition and in the Electronic Registry of suppliers of Mer-link, as well as that the attorney is subject to the laws and courts of Costa Rica regarding all acts or contracts entered into or they must be executed in the country and expressly renounces the laws of their domicile. (Article 232 Commercial Code.)

Mr. NAME OF AUTHORIZED, is entitled to on our behalf and representation can receive, desist from the request, notify, replace, resume and interpose all actions and resources necessary for the best exercise of this power.

City, Date
FIRM

____
APPOINTMENT NAME / ID DATA
I accept power:

____
AUTHORIZED NAME / ID DATA

Text 4: National Register of Companies

As of publication, the interested party has 30 working days to exercise its opposition in the Judicial Branch.

All information must be corroborated in La Gaceta and the corresponding scope.

Method for consultation:

The simplest way to locate a company is by using the "SEARCH" key and typing the consecutive number of the legal identity card.

It is not necessary to indicate the numbers corresponding to entity type, namely 3-101- 3-012- etc. only the consecutive one.

When the consecutive has less than 6 digits, you should not type zeros to the left, only the number that composes it. Axis: 3-101-001234 (only 1234 must be entered).

Text 5: Services

In this registry they are registered: individual companies of limited responsibility, companies in collective name, limited partnerships, limited liability companies, corporations; companies governed by special laws, stock exchanges, stock exchange positions, investment companies, investment fund companies, securities companies, clearing and settlement companies, risk rating agencies, supplementary pension fund operators, financial companies of a non-banking nature, private banks, labor corporations, insurance trading companies and sports corporations.

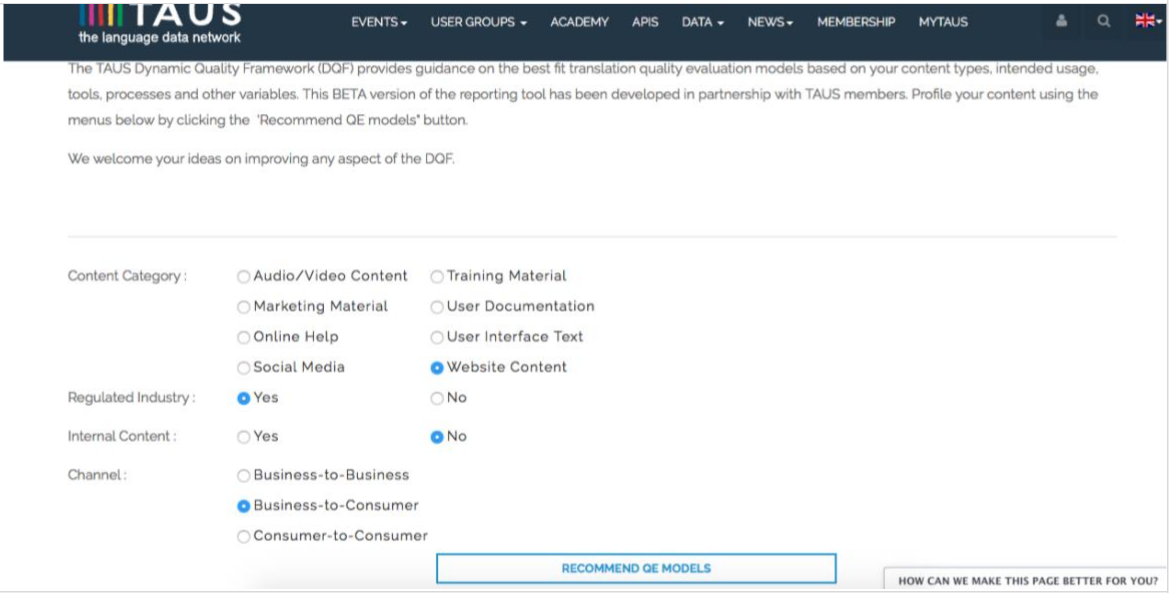
Registration of foundations; professional activity societies; civil and sports associations. Registration of reforms to the legal entities indicated; also inscription of powers, granted by natural persons; conferred by national societies; granted abroad and conferred by foreign companies; Opening of branches of foreign companies.

Also, insolvency is inscribed; insanity tutorial; executorships; sworn corridors; reservation of name and, in addition, supervision of civil associations; as well as information consultation in the automated system and in the traditional tomes.

Appendix 17. Rejected Words Report of System 2

File name	Total Words	Rejected Words	Percentage
DGT_es_en_06.tmx.tmx.noprop.tmx	2.093.516	198.212	9%
ALIGNED_es_en.tmx.tmx.noprop.tmx	3.313.263	151.960	5%
ECB_es_en.tmx.tmx.noprop.tmx	3.514.411	652.527	19%
News_es_en.tmx.tmx.noprop.tmx	6.286.578	96.093	2%
DGT_es_en_01.tmx.tmx.noprop.tmx	7.832.778	737.678	9%
DGT_es_en_05.tmx.tmx.noprop.tmx	8.839.176	704.228	8%
DGT_es_en_03.tmx.tmx.noprop.tmx	9.399.880	762.876	8%
DGT_es_en_04.tmx.tmx.noprop.tmx	9.836.761	809.472	8%
DGT_es_en_02.tmx.tmx.noprop.tmx	10.316.145	934.136	9%
		Average of rejected words	9%

Appendix 18. TAUS FORM to Find the Evaluation that Best Suits Our MT Output Evaluation



The TAUS Dynamic Quality Framework (DQF) provides guidance on the best fit translation quality evaluation models based on your content types, intended usage, tools, processes and other variables. This BETA version of the reporting tool has been developed in partnership with TAUS members. Profile your content using the menus below by clicking the 'Recommend QE models' button.

We welcome your ideas on improving any aspect of the DQF.

Content Category :	○ Audio/Video Content	○ Training Material
	○ Marketing Material	○ User Documentation
	○ Online Help	○ User Interface Text
	○ Social Media	● Website Content
Regulated Industry :	● Yes	○ No
Internal Content :	○ Yes	● No
Channel :	○ Business-to-Business	
	● Business-to-Consumer	
	○ Consumer-to-Consumer	

Appendix 19. Golden Standard Translation

Text 1: Frequently Asked Questions

3. What are the main advantages of registering in CrearEmpresa for entrepreneurs?

CrearEmpresa aims to support those interested entrepreneurs in the creation of their business. It also attempts to offer the State services in a fast and simple fashion.

Through a single electronic window, the interested applicants would be able to carry out all the necessary formalities online in order to establish and start operating their company or business.

The following are the different procedures that CrearEmpresa offers:

a. Registration of corporations in the National Registry.
b. Electronic legalization of social books in the National Registry.
c. Publication of edicts in the official journal La Gaceta.
d. Certification of legal identity by the National Registry.
e. Certification of real estate property.
f. Certification of digitized cadastral plans.
g. Certificate of land use right.
h. Issuance of the sanitary permit to operate, by the Ministry of Health.
i. Issuance of the Enviromental Feasibility, by the National Environmental Technical Secretariat (SETENA) for businesses of low environmental impact.
j. Veterinary Certificate, issued by the National Service of Animal Health (SENASA) of the Ministry of Agriculture and Cattle.
k. Verification of the insurance policy for occupational risk, by the National insurance Institute (INS).
l. Patents.
m. Registration of employers in the Costa Rican Social Security Administration S.A.
n. Registration of taxpayers in the Tax Authority.

5. What does registering a business mean?

The registration of a business refers to the creation of a new and legal society within the National Registry. Only notaries have the legal authorization to complete this procedure.

Text 2: Simplified Procedure for Foreigners

REQUIREMENT B

[Date]

Mr.
Chief Executive Officer
General Department of Migration and Immigration of Costa Rica

Dear Sir,

The company [official name of the company] requests a residence permit as a business visitor, on behalf of Mr. or Mrs. [name], of legal age, [marital status], business agent, resident of [address from its country of origen], temporarily residing in [address in Costa Rica], citizen [nationality], holder of passport number [passport number], who is interested in making businesses in Costa Rica for a year on behalf of [official name of the company], as welll as registering as a provider and participating in public tenders through the public procurement system, Mer-link.

Likewise, this person shall not accrue the payment of wages or fees within the country. This request in accordance with article 88, section 2) of the General Law of Migration and Aliens 8764, effective as of March 1st, 2010, complying with all requirements.

Sincerely,
[Full name]
[Position: Chief Executive, Legal Representative]
[Company]

NAME OF REPRESENTATIVE/ID INFORMATION

I accept the power of attorney

Text 3: National Registry of societies

From the date of publication, the interested party has 30 business days to claim its opposition in court.

All information must be supported in the corresponding official journals, La Gaceta and Alcance.

Consultation system:

The simplest way to locate a company is by typing the “Search” key and typying the consecutive identification number of the corporation.

It is not necessary to indicate the number of the type of company, namely 3-101-3-012- etc., only the consecutive one.

When the consecutive number has fewer than 6 digits, you should not type the zeros to the left, but only the numbers. For example: 3-101-001234 (only type the numbers 1234).

Text 4: Services

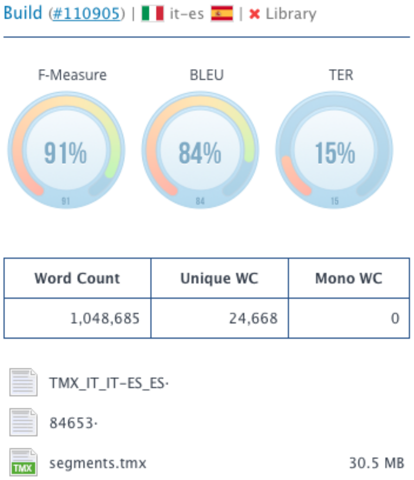
In this registration, you can register the following: single-owner limited liability companies, partnerships, limited partnerships, limited liability companies, public limited companies; companies governed by special laws, stock exchange, exchange posts, investment trusts, investment fund operators, central security depositories, clearing and settlement companies, risk assessing companies, operating companies of complementary pension funds, financial companies of non-banking nature, private banks, public limited labour companies, insurance trading companies and sports corporations.

Registration of endowments, professional associations, civil and sports associations.

Registration of reforms to legal entities, of powers of attorney, granted by individual persons, conferred by local companies, granted overseas and conferred by foreign companies; opening of branch offices of foreign companies.

Likewise, you can register cases of insolvency, insanity, tutoring, executorships, brokers, reservation of a trading name and audits of civil associations. You can also find more information in the automated system and in the corresponding volumes.

Appendix 20. KantanMT Test: A Story of Success



Build (#110905) | it-es | ✖ Library

Word Count	Unique WC	Mono WC
1,048,685	24,668	0

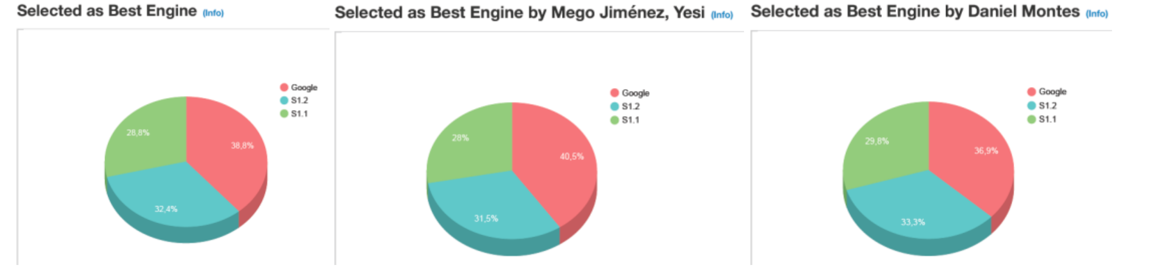
TMX_IT_IT-ES_ES·

84653·

segments.tmx 30.5 MB

Prior to training an engine using the corresponding datasets for this research paper, I carried out several trainings with small datasets. In this case, I compiled the TMX files of a client that works in the clothing industry, and its website content had been localized. These translation memories belong to a translation agency; thus, they are professional translations that have been already checked and delivered to the client. The training of 1,048,685 words which corresponds to a translation memory of 32 MB successfully scored 15% of TER, meaning that the post-editing is minimal. In fact, after a few translation jobs, upon unseen strings, the output text requires minimal corrections and the text is more than ready for dissemination. This small experiment demonstrates that 1) there is no need for large amounts of datasets, not even a monolingual data set was required to get such satisfactory results; 2) it also shows that experienced translators (who have their own TMs) would benefit the most out of this technology by making minimal effort; 3) automated metrics highly correlate with the human quality expectations.

Appendix 21. Ranking Evaluation Results



Selected as Best Engine (Info)

Selected as Best Engine by Mego Jiménez, Yesi (Info)

Selected as Best Engine by Daniel Montes (Info)