Evaluation of Machine Translation models for French - Romanian

Raluca Andreea Gînga
raluca.ginga@s.unibuc.ro

Diana Ionescu
diana.ionescu2@s.unibuc.ro

Abstract

This report has the main purpose to compare and evaluate a variety of machine translation models that are able to directly translate sentences from French to Romanian. This specific pair of languages was not approached yet in the literature, giving us the possibility to explore and experiment with different models, varying from Simple Deep Learning architectures to Pretrained Models such as Marian MT (Helsinki NLP) using OPUS Machine Translation(Tiedemann and Thottingal, 2020).

1 Introduction

Neural Machine Translation is a relatively new approach to Machine Translation. In general, the networks proposed have an encoder-decoder structure and they have a better performance than the methods of Statistical Machine Translations (Bahdanau et al., 2014), being currently state-of-the-art technology in this domain.

In this paper, we present a few neural machine translation systems, some of them are pretrained models (for example: MarianMT) and the others are trained from scratch for the French-Romanian pair languages.

The main idea of a machine translation system is that, given a sequence in a source language (in our case, French) we use a neural network to predict the probable words to the target language (Romanian). It is good to know that, for each language, we apply some preprocessing steps and map each token to a number, in order to create a source and a target vocabulary.

The evaluation of the models are both, automatic (using the BLEU score, TER score and Meteor score) and made by us, as Romanian native speakers and French speakers. The human evaluation is made based on the adequacy and fluency methods.

There are many neural machine translation models, however we have not found models to implement the direct translation from French to Romanian, yet, we should mention that the French language was used as pivot for the English-Romanian translation for an unsupervised model in the paper (Li et al., 2020). However, the only reference that we've found in the "literature" is in Helsinki NLP webpage regarding the model from French to Romanian on Tatoeba databaset with a BLEU score of 42.1 using Helsinki NLP pretrained model.

The present task has the main purpose to use Machine Translation models in order to evaluate their performance on French - Romanian language pair. The remainder of the paper is organized as follows: section 2 is describing the dataset and the data analysis and preprocessing made to the dataset in order to understand it better. Section 3 describes the details of the implemented Deep Learning techniques and other methods that were tried, but didn't bring any good results. Section 4 is discussing about the evaluation metrics that we used on our datasets: both human evaluation for a demo sample of 50 sentences from the test dataset and automatic evaluation used for both validation and testing datasets. In 5, there are discussed the result bought by each of the successful approaches, followed by a little demo on test phrases in section number 6. The last section (7) concludes the paper about the implemented methods and comes with some further work that could be done in order to research more about this task.

2 Background

For this specific task, we used WikiMatrix v1 found on https://opus.nlpl.eu/WikiMatrix.php website. WikiMatrix (Schwenk et al., 2019) is one of the largest and complete extraction of different kinds of sentences present on Wikimedia across multiple languages. This corpus is used by a lot of NLP researchers for training, evaluation and comparing new translation models or other multilingual models purposes.

A sample of the dataset is in the figure from

below (1), where we can clearly see that we have a total number of 206k of entries, 90 of them being unstranslated.

Dumnezeu călăuzește către lumina Sa pe cine voiește.
 = Allah guide vers Sa lumière qui Il veut.

Milostivenia Domnului tău este mai bună decât ceea ce ei adună.
 = Cependant, les bonnes œuvres qui persistent ont auprès de ton Seigneur une meilleure récompense et une belle espérance.

Dumnezeu este Văzător a ceea ce fâptuiesc.
 = Voilà comme il regarde, et ce qu'il regarde ! ».

Deasupra lor sunt trei îngeri, unul dintre ei purtând o ramură de palmier.
 = Au-dessus d'eux sont trois anges, l'un portant une palme.

Dacă vă este teamă că nu veți fi drepți cu ele, luați-vă o singură femeie ori pe cele stăpânite de dreapta voastră.
 = Mais, si vous craignez de n'être pas équitable avec celles-ci, alors une seule, ou des esclaves que vous possédez.

Number of sentences in Romanian corpus = 206444
Number of sentences in French Corpus = 206444
Untranslated sentences in the dataset = 90
Phrases that contain more than one sentences in Romanian 5382 and in French 4667

Figure 1: Sample of the dataset

2.1 Data Analysis & Preprocessing

Because we wanted to get to know better the dataset and have a general overview about it, we decided to first analyze the words length. We used box-plot in order to see better if we're dealing with outlier and trying to detect the first problems that can occur in a Machine Learning project. As we can notice in figure 2, we have a lot of outliers cocerning the Romanian sentences length (and it is quite obvious we have French sentences outliers as well), but first, we decided to drop the outliers from both Romanian and French sentences using Interquartilic Range (IQR). In order to detect the outliers, we computed the first and third quartile of our dataset sentences, computing then the interquartilic range (given by the difference between the third and the first quartile), proceeding with the index identification of the entries that are not belonging to the outliers range $[Q_1 - 1.5 * IQR, Q_3 + 1.5 \times IQR]$, where Q_1 is the first quartile, Q_3 is the third quartile and IQR is the interquartilic range.

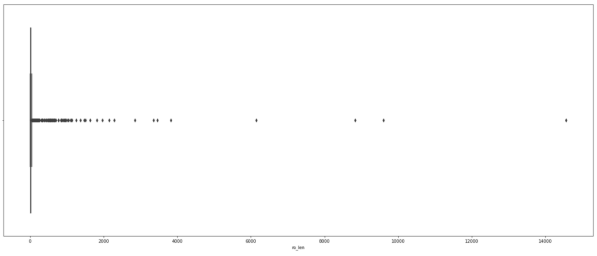

Figure 2: Words length of the dataset in box-plot

We detected a total of 9561 outliers, 5827 from them being given by the Romanian sentences and 3734 outliers by the French sentences (after doing the outlier removal for Romanian outliers). After doing the removal, we got a total of 196791 samples. We can see the distribution of the words length present in Romanian and French datasets in the two figures from below.

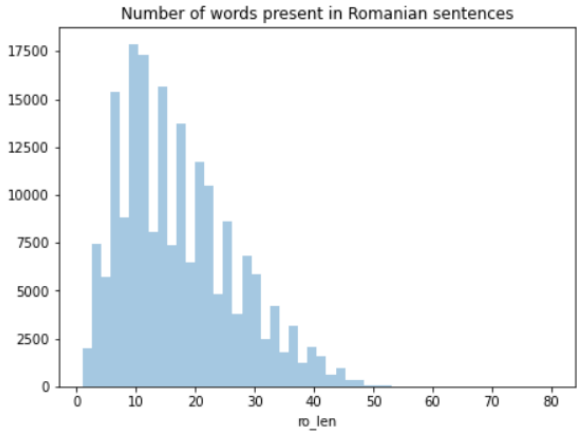

Figure 3: Words distribution for Romanian sentences

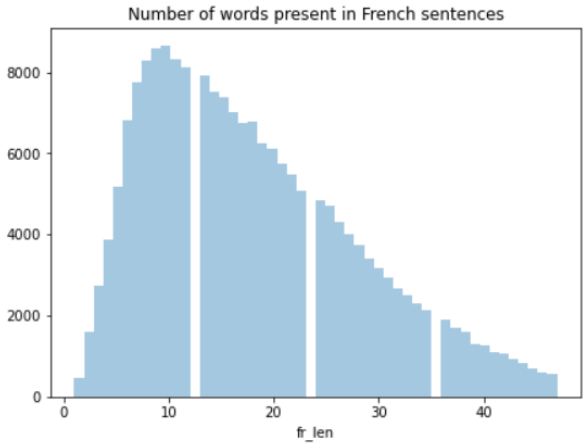

Figure 4: Words distribution for French sentences

3 System Overview

3.1 Dataset

Because we noticed that the training time on the whole corpus is taking a lot of time exceeding our laptops and even Google Colab's GPU capacity, we decided to take only a small sample of approximately 2% of the dataset, obtaining a dataframe of approximately 3000 sentences. Then, we did train-test-split on the dataset in 70-26.5-3.5 form.

After doing the train-test-split, we wanted to see the most frequent words from our dataset. In order to generate the word cloud and to have a better representations of the most common words, we applied some preprocessing steps like unidecoding and obtaining the non-diacritical form of the sentences, clearing the special characters, lowercase of the words, tokenizing, stopwords removal and words shorter than 3 characters removal.

In the figure from below, we can see the words distribution for Romanian and French sentences.

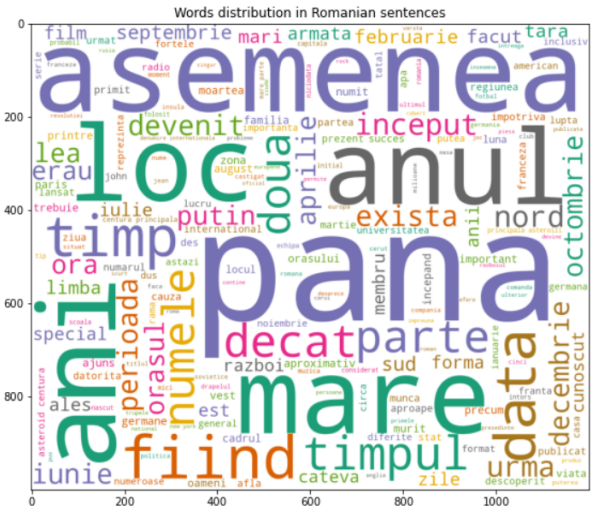

Figure 5: Words distribution for Romanian sentences

Figure 6: Words distribution for French sentences

3.2 Deep Learning techniques

For this project we tried a large variety of algorithms, providing us both very poor results, but good results as well. In the "Methods tried but not successful" subsection, there are described some algorithms tried using Pytorch, but not giving very good results. However, we decided to include that subsection too because there were applied different methods and we selected the methods that worked best for our problem.

1. Simple Recurrent Neural Network (SimpleRNN)

   In a Recurrent Neural Network the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously. However, this simple RNN has short-term memory.

   As data preparation step, we used Tokenizer in order to obtain the numeric representations of the words, sequence padding to the maximum length of the Romanian sentence, obtaining a French vocabulary size of 13k words and a Romanian vocabulary size of 14k words.

   As model, we used a Simple RNN with 128 neurons, taking into account the embedding formed by the french vocabulary size and an embedding size of 128, resulting into an output sequence length of 64 words. Then, we used a Time Distributed dense layer of Romanian vocabulary size in order to get the outputs. The loss function was 'sparse categorical crossentropy', we used an Adam optimizer with 0.01 as learning rate and the main metric used was accuracy. We trained this model for 30 epochs and the time per epoch on Google Colab was approximately 7 seconds.

2. Gated Recurrent Unit (GRU)

   In sequence modeling techniques, Gated Recurrent Unit is the newest entrant after RNN and LSTM, offering some good improvements over the other two mentioned algorithms. GRU is an advancement of the standard Recurrent Neural Network, being introduced by (Chung et al., 2014).

   As data preparation step, we used Tokenizer in order to obtain the numeric representations of the words, sequence padding to the maximum length of the Romanian sentence, obtaining a French vocabulary size of 13k words and a Romanian vocabulary size of 14k words.

   As model, we used a GRU with 128 neurons. We trained this model for 30 epochs and the time per epoch on Google Colab was approximately 5 seconds.

3. Pretrained Models

   - OPEN MT based on Marian MT
     Marian MT represents a transformer with an encoder-decoder of 6 layers each, being mainly developed by the Microsoft Translator team.
     For this, we used OpenNMT with "opus-mt-fr-ro" model. We chose this technique from Hugging Face because it's

already trained and in order to see how it's performing against the real translation provided in the dataset.

- mBART-50

  mBART-50 represents a multilingual Seq2Seq pretrained model (Tang et al., 2020). This model was created and introduced in order to showcase the fact that multilingual neural machine translation models could be created through fine-tuning. But, instead of fine-tuning only on a single direction, this model is fine-tuned on a lot of directions at the same time.

  In our case, we used mBART-50 One to many multilingual machine translation model. In this model, we specified the source and the target languages.

- Translate module

  Translate is a simple and powerful translation module written in python that had the main purpose on making translations easier for everyone without so much effort. In this regard, we used this tool in order to see how automated APIs perform on our dataset and to see the final result.

3.3 Methods tried but not successful

In this section, we can include a two-layers LSTM model that takes as input the one hot encoding representation of the source sequences (it has the shape (batch_size, max_length_sequence, vocabulary_length)). The output is an array that contains the predictions of shape (batch_size, max_length_sequence). More precisely, the architecture contains one embedding layer for obtaining word embeddings, then we apply two LSTM layers and, the output is fed into a linear layer that will transform the input into an array of batch_size elements with max_length_sequence features; finally, an Softmax activation function.

The optimizer that we have used is Adam with different learning rates (0.01, 0.001) and the loss function is CrossEntropyLoss.

After running for different number of epochs (3, 50, 100, 200), we have observed that the output of the model (the translations) contains only the words that have a small index in the vocabulary (for example: 'de de de a de de de') and the BLEU score is zero.

Model	BLEU	METEOR	TER
SimpleRNN	0.66	0.44	1.04
GRU	0.8	0.52	1.1
Translate	0.875	0.54	2.23
OpusMT	**0.86**	**0.66**	**0.72**
mBART-50	0.4	0.57	0.93

Table 1: Machine Translation scores on Validation

The model was trained on a GPU and the train on one epoch takes approximately 40 seconds.

4 Evaluation metrics

4.1 Human evaluation metrics

Fluency. This metric is used to measure the fluency of a sequence, without taking into consideration the meaning of the source sequence (Snover et al., 2009).

Adequacy. This metric measures whether the translation is correct with respect to the meaning of the source sequence, even if the fluency of the translation is poor (Snover et al., 2009).

We should also mention the fact that, the assessors will grade a translation from 1 to 5, where 1 is wrong and 5 is correct.

4.2 Automatic evaluation metrics

BLEU score is a quality metric computes the overlaps n-grams (from one to four) between the target sequence and translated sequence (Bojar, 2015) and it is one of the most commonly used.

METEOR is a metric that computes the similarity between the translated sequence and target sequence by counting the exact word matches and, for the words that do not match, a stemmer is applied. In case of reordering the words, some penalties are added (Snover et al., 2006).

TER or translated error rate measures the amount of data that has to be modified, in the translated sequence, in order to correspond to the target seqence (Snover et al., 2006).

5 Results

In tables 1 and 2, it can be seen that the best performing model in terms of BLEU score, METEOR score and TER is OPUS Machine Translation for French-Romanian corpus. In terms of translation error rate, again, OPUS MNT seems to do the greatest job followed by mBART-50 (although giving poor results on BLEU score).

Model	BLEU	METEOR	TER
SimpleRNN	0.58	0.35	1.1
GRU	0.71	0.56	1.14
Translate	0.93	0.54	2.21
OpusMT	**0.96**	**0.67**	**0.69**
mBART-50	0.5	0.6	0.89

Table 2: Machine Translation scores on Test

6 Demo on test phrases

In this demo, we wanted to manually evaluate a set of 50 phrases present in the test dataset and compare our results.

We extracted 50 phrases from the test dataset and provided the scores in a shared Excel of the adequacy and fluency of the translated phrases. During this process, we encountered some different problems:

- phrases that are correctly translated by the model, but those "actual" phrases were misleading
- "actual and real" Romanian phrases contain a lot of redundant information comparing to the French phrases and it is obvious this redundant information (unnecessary details) influenced the BLEU score of those phrases as well
- misleading French and Romanian phrases
- some actual Romanian phrases that are not correct and contain wrong information in relationship with the French phrases

French	Actual Translation	Predicted Translation
En linguistique la notion de connecteur n ' a pas de définition généralement acceptée et ce que l ' on considère qu ' elle représente n ' est pas nommé avec ce terme par tous les auteurs	In lingvistica notiunea de conector nu are o definitie general acceptata si nici ceea ce se considera ca reprezinta nu este denumit de toti autorii cu acest termen	In limba notiunea de conector nu are o definitie general acceptata si ceea ce se considera a fi "n" este numit cu acest termen de catre toti autorii
Ce psaume est également utilisé lors des célébrations pénitentielles comportant le sacrement de pénitence et de réconciliation	Acest Psalm este de asemenea utilizat in timpul festivitatilor penitentiale inclusiv in Sacramentul Pocaintei si Reconcilierii	Acest psalm este folosit, de asemenea, la sarbatorile penitentiale care implica sfintirea penitentei si a reconcilierii.
Années 1930	In anii 1930	Anii 1930
Celui-ci l'emporte au second tour avec plus de 52 des voix	Acesta a castigat al doilea tur cu peste 52 din voturi	Acesta castiga in a doua runda cu mai mult de 52 de voturi.
À cette époque elle travaille également en Israël	El a continuat aceasta activitate si in Israel	Pe vremea aceea, ea lucreaza si in Israel.

Figure 7: Sample of phrases. Approach for evaluating 50 test phrases

Adequacy_Raluca	Fluency_Raluca	Adequacy_Diana	Fluency_Diana	Observatii despre traduceri
3	4	2	4	
5	5	5	5	
5	5	5	5	traducerea romaneasca e misleading si influenteaza negativ scorul pentru predicted
5	5	5	5	
4	4	5	4	traducerea tinta este usor diferita de varianta corecta ('El' in loc de 'Ea')

Figure 8: Scores and observations samples of the phrases. Approach for evaluating 50 test phrases

Person	Adequacy	Fluency	BLEU
Raluca	4.36	4.34	0.92
Diana	4.46	4.6	0.92

Table 3: Human evaluation along with BLEU score used for this sample of phrases. Adequacy1 and Fluency1 are the mean scores of Raluca. Adequacy2 and Fluency2 are the mean scores of Diana

7 Conclusion and Further Work

In the current project, it was solved the task related to the evaluation of different Neural Machine Translation models for French-Romanian language pair. There were applied various methods, including the application of basic Neural Network models to pretrained models like Open NMT by Helsinki NLP based on Marian MT, mBART-50 and a module called Translate that is based on Google Translate API. The enormous potential was given by Marian MT model, giving a BLEU score of 0.86 on validation dataset and 0.96 on test dataset. Then, we decided to also evaluate 50 of the phrases in a manual form by giving scores in terms of adequacy and fluency. As we could clearly see, there are discrepances between the scores provided by Raluca and the scores computed by Diana and this thing is normal, since human factor is subjective.
As far as other improvements are concerned:

- some extra tests could be run in order to understand and correct the models that do not return any correct translation, since we have already noticed some mistranslations in the Romanian dataset and the first modification would be to correct them
- taking into consideration that there are not many papers concerning the French - Romanian translation domain, it will be interesting to train and test the models for different dialects, for example from Moldavian to French
- Using more powerful GPUs to train on the entire corpus and trying transformers or other models

.

7.1 Contributions

- Raluca was focused on doing the dataset preprocessing and statistical data analysis in order to understand it better. She wrote the code for all of the successful algorithms that work for our problem: Simple RNN, GRU and pretrained models, computed some functions for the automatic machine translation metrics like BLEU, Meteor and TER scores. Contribution to paper: Abstract, Introduction, Background section, System Overview section (3) without the "Methods tried but not successful" subsection, Results, Demo on test phrases and Conclusion

- Diana wrote code for the methods that didn't bring so many good results and she was focused on trial-error models, knowing further in which models to invest time and energy. She evaluated the Translate module against the test dataset. She was also involved in Human Evaluation of the 50 test phrases. Contribution made to paper are: introduction, methods tried but not successful subsection from section 3, section 4 and the final section that concludes the paper 7

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondřej Bojar. 2015. Automatic mt evaluation.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020. Reference language based unsupervised neural machine translation.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.