WMT Shared Task: Similar Language Translation For Romance Languages: Spanish → Romanian

1st Semester of 2021-2022

Adrian Răzvan Iordache
adrian.iordache@s.unibuc.ro

Andrei-Cristian Gîdea
andrei.gidea@s.unibuc.ro

Abstract

Within the MT and NLP communities, English is by far the most resource-rich language. MT systems are most often trained to translate texts from and to English or they use English as a pivot language to translate between resource-poorer languages. Thus, we chose to build a system for the automatic translation of sentences from Spanish into Romanian. The aim of this project is largely based on understanding the algorithms that produced SOTA in the field of NLP, thus being a project that represents an end-to-end implementation of these algorithms, with the presentation of the results obtained and their comparison with other frameworks.

1 Introduction

Machine translation (MT) works as an interface that handles language ambiguity concerns via automatic translation between two different languages. Neural machine translation (NMT) attains state of the art (SOTA) results for both high and low-resource language pairs translation ((Cho et al., 2014); (Luong et al., 2015); (Sutskever et al., 2014)). The NMT utilizes an artificial neural network to predicts the likelihood of a sequence of words. But NMT requires a sizeable parallel corpus to get effective MT output, challenging for low-resource pair translation.

With the widespread use of MT technology, there is more and more interest in training systems to translate between languages other than English. One evidence of this is the need of directly translating between pairs of similar languages. The main challenge here is how to take advantage of the similarity between languages to overcome the limitation given the low amount of available parallel data to produce an accurate output.

As part of the similar language translation's subtask for Romance Languages, namely Spanish (ES) and Romanian (RO), we have attempted to build Neural Machine Translation (NMT) models using the Pytorch framework (Paszke et al., 2019), thus trying to build an end-to-end NMT system, without the help of any special toolkit created for this task, in order to better understand the algorithms behind it. In the project we wanted to try several variations of the models, implementing both a system based on RNNs, but which turned out to be much too slow for what we wanted, and a system based on a Transformer architecture. Among the main experiments, the impact of the size of the vocabulary used by the model on the metrics was analyzed, as well as the variation of the parameters that constitute the model. In all experiments, the data were transformed into lowercase and tokenized using byte-pair encoding (BPE) using the youtokentome [1] library for our framework and in JoeyNMT, the data were tokenized using subword-nmt library and not lowercased as before.

All the code used in this project can be found in this Github repository[2].

Section 2 presents the latest discoveries in the field, Section 3 presents the data set used, the number of sentences found in each source as well as the preprocessing applied to it. Section 4 describes the data cleaning methods. Section 5 presents the methods and models used, as well as the theory behind them. Section 6 describes the evaluation methods of the obtained systems, Section 7 emphasizes the contribution of each of us and Section 8 presents the conclusions we have reached working on this project and what could have been improved.

2 Related work

In recent times, there has been an increase of research interest in low-resource MT scenarios ((Jawahar et al., 2021); (Baziotis et al., 2020)). NMT models, specifically those based on the Transformer architecture, have been shown to perform

[1] https://github.com/VKCOM/YouTokenToMe
[2] https://github.com/AdrianIordache/Machine-Translation-ES-RO)

well when translating between similar languages ((Przystupa and Abdul-Mageed, 2019); (Adebara et al., 2020)), low resource scenarios (Adebara et al., 2021), and in contexts not involving English (Fan et al., 2020).

3 Data

For the training of the NMT systems sentence aligned parallel corpora are required. We used all parallel corpora suggested by the SLT shared task organizers. For Spanish-Romanian the following parallel corpora were used:

1. Wiki Titles v3 (253,770 sentence pairs)
2. Europarl v8 (387,653 sentence pairs) (Koehn, 2004)
3. Tilde MODEL (3,770 sentence pairs) (Rozis and Skadiņš, 2017).
4. JRC-Acquis (451,849 sentence pairs) (Steinberger et al., 2006)

3.1 Pre-processing

For simplicity and because the purpose of the project was not to obtain the best possible results, being an aspect to be modified in the future probably, we chose to train the models on the lowercase data.

But we did not feed the model with the data as it comes, but we chose to tokenize it by the byte-pair encoding (BPE) (Sennrich et al., 2015) algorithm trained on the train data.

BPE is a simple form of data compression algorithm in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur in that data.

4 Data cleaning and Analysis

We perform data cleaning on the ES-RO language pair. For cleaning, we run the langid tool (Lui and Baldwin, 2012) on the concatenation of the source and target and remove sentences that are not identified as belonging to Romanian language. In Table 1, we provide some examples of data points we remove from the training data during data cleaning. These examples are removed because the claimed language is different from the language predicted by langid. After cleaning, we are left with $\sim$ 840k clean sentences out of $\sim$ 1M sentences for the Spanish and Romanian pair. We note that removed data comprise large portions of the dataset, thus confirming our concerns about data quality.

Sentence	Claimed	Predicted
(La séance, suspendue à 11h35)	Romanian	French
(Die Sitzung wird.)	Romanian	Deutsch
(La seduta, sospesa)	Romanian	Italian
(Se levanta la sesión)	Romanian	Spanish
(For the results and other details.)	Romanian	English

Table 1: Examples removed from our training data. "Claimed" refers to the expected language as coming from source, while "predicted" is what langid.py identified.

Sentence	Claimed	Predicted
Cu toate acestea, Václav Klaus.	Romanian	Deutsch
Summitul G-20 de la Pittsburgh	Romanian	Spanish
Acest lucru ar prelungi.	Romanian	Swedish

Table 2: The examples that Langid wrongly classifies as being from a language other than Romanian.

After that, we decided to investigate if there were duplicated examples in the dataset, and our suspicions were confirmed. Thus, we chose to drop the duplicate sentences in the data set and keep only unique examples, thus remaining with $\sim$ 700k sentences to train on.

5 Models

To evaluate the importance of different components of the Transformer, we varied our base model in different ways, measuring the change in performance on Spanish-Romanian translation on the development set.

5.1 Transformer from scratch

The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. The Transformer was proposed in the paper (Vaswani et al., 2017).

The transformer architecture and the seq2seq model with attention are similar to each other in overall: the source sequence embeddings are fed into n repeated blocks. The outputs of the encoder's last block are then used as attention memory for the decoder. The target sequence embeddings is similarly fed into n repeated blocks in the decoder, and the final outputs are obtained by applying a linear layer with vocabulary size to the last block's outputs.

It can also be seen that the transformer differs to the seq2seq with attention model in three major places:

1. A recurrent layer in seq2seq is replaced with a transformer block. This block contains a self-attention layer (muti-head attention) and a network with two linear layers (position-wise FFN) for the encoder. For the decoder, one more multi-head attention layer is used to take the encoder state.

2. The encoder state is passed to every transformer block in the decoder, instead of using as an additional input of the first recurrent layer in seq2seq.

3. Since the self-attention layer does not distinguish the item order in a sequence, a positional encoding layer is used to add sequential information into each sequence item.

5.1.1 Token Embedding

The first step is feeding out input into a word embedding layer. A word embedding layer can be thought of as a lookup table to grab a learned vector representation of each word. Neural networks learn through numbers so each word maps to a vector with continuous values to represent that word.

5.1.2 Positional Encoding

In order for the model to make sense of a sentence, it needs to know two things about each word: what does the word mean? And what is its position in the sentence?

The embedding vector for each word will learn the meaning, so now we need to input something that tells the network about the word's position.

As each word in a sentence simultaneously flows through the Transformer's encoder/decoder stack, the model itself doesn't have any sense of position/order for each word. Because the transformer encoder has no recurrence like recurrent neural networks, we must add some information about the positions into the input embeddings. This is done using positional encoding.

5.1.3 Encoder

The Encoders layers job is to map all input sequences into an abstract continuous representation that holds the learned information for that entire sequence. It contains 2 sub-modules, multi-headed attention, followed by a a feed-forward neural network. There are also residual connections around each of the two sublayers followed by a layer normalization.

Multi Head Attention

Multi-headed attention in the encoder applies a specific attention mechanism called self-attention. Self-attention allows the models to associate each word in the input, to other words.

Residual Connections & Layer Normalization

The multi-headed attention output vector is added to the original positional input embedding. This is called a residual connection. The output of the residual connection goes through a layer normalization.

Position-Wise Feed Forward Network

The normalized residual output gets projected through a pointwise feed-forward network for further processing. The pointwise feed-forward network is a couple of linear layers with a ReLU activation in between. The output of that is then again added to the input of the pointwise feed-forward network and further normalized.

The residual connections help the network to train, by allowing gradients to flow through the networks directly. The layer normalizations are used to stabilize the network which results in substantially reducing the training time necessary. The pointwise feedforward layer is used to project the attention outputs potentially giving it a richer representation.

5.1.4 Decoder

The decoder's job is to generate text sequences. The decoder has a similar sub-layer as the encoder. It has two multi-headed attention layers, a pointwise feed-forward layer, and residual connections, and layer normalization after each sub-layer. These sub-layers behave similarly to the layers in the encoder but each multi-headed attention layer has a different job.

The decoder is autoregressive, it begins with a start token, and it takes in a list of previous outputs as inputs, as well as the encoder outputs that contain the attention information from the input. The decoder stops decoding when it generates a token as an output.

First Multi Head Attention

This multi-headed attention layer operates slightly different. Since the decoder is autoregressive and generates the sequence word by word, we need to prevent it from conditioning to future tokens.

Second Multi Head Attention

For this layer, the encoder's outputs are the queries and the keys, and the first multi-headed

attention layer outputs are the values. This process matches the encoder's input to the decoder's input, allowing the decoder to decide which encoder input is relevant to put a focus on. The output of the second multi-headed attention goes through a pointwise feedforward layer for further processing, identical to the one in the encoder layer.

5.2 Comparison with other framework

Just for a brief comparison, we wanted to see what the difference would be between our implementation and the implementation of an end-to-end system in a framework, more precisely JoeyNMT (Kreutzer et al., 2019). Table 3 reflects the difference in score.

6 Evaluation

We evaluated our models on both the Dev and Test sets. We used the checkpoint with the best BLEU score as evaluated on DEV as our best model. We used a beam size of four during evaluation on both Dev and Test and evaluated on de-tokenized data.

Both the BLEU score, the TER score and the chrF++ score are calculated using the sacrebleu (Post, 2018) library.

6.1 BLEU Score

The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence introduced in (Papineni et al., 2002). The BLEU metric scores a translation on a scale of 0 to 1, in an attempt to measure the adequacy[3] and fluency[4] of the MT output. The closer to 1 the test sentences score, the more overlap there is with their human reference translations and thus, the better the system is deemed to be.

6.2 TER Score

The TER score (Snover et al., 2006) measures the amount of editing that a translator would have to perform to change a translation so it exactly matches a reference translation. By repeating this analysis on a large number of sample translations, it is possible to estimate the post-editing effort required for a project.

[3] **Adequacy:** Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

[4] **Fluency:** Is the output good fluent Romanian? This involves both grammatical correctness and idiomatic word choices.

6.3 chrF Score

The chrF score (Popović, 2015), or character-level F-score, is a simple but very effective metric. Informally, it measures the amount of overlap of short sequences of characters (n-grams) between the MT output and the reference.

6.4 Beam Search

Beam Search makes two improvements over Greedy Search:

1. With Greedy Search, we took just the single best word at each position. In contrast, Beam Search expands this and takes the best N(four in our case) words.

2. With Greedy Search, we consider each position in isolation. Once we had identified the best word for that position, we did not examine what came before it (ie. in the previous position), or after it. In contrast, Beam Search picks the N(four in this case) best sequences so far and considers the probabilities of the combination of all of the preceding words along with the word in the current position.

Evaluation on Dev set. We show sample outputs from our variations of models (from Dev data) in Table 5.

Evaluation on Test set. We show sample outputs from our variations of models (from Dev data) in Table 6.

Our Test set performance is evaluated using BLEU (Papineni et al., 2002), TER (Snover et al., 2006) and chrF (Popović, 2015). We report the scores in Table 8.

7 Contributions

Iordache Adrian

1. Designing framework architecture

2. Writing the code for distributed training on multiple GPUs with automated logging system for end-to-end framework

3. Writing the code for inference/translation using k-beam search

4. Researching and writing code for the Transformer architecture with multi-headed attention layers for the end-to-end translation system

	vocab_size	N	d_model	dff	h	d_k	d_v	P_drop	train steps	VAL Set		TEST Set	
										BLEU	TER	BLEU	TER
Our System	4k	6	512	2048	8	64	64	0.1	20k	23	0.766	22.9	0.775
JoeyNMT	4k	6	512	2048	8	64	64	0.1	20k	27.2	0.705	27.2	0.697

Table 3: Difference between our system and JoeyNMT system

Sentence	**Claimed**	**Tool**	**Predicted**
Acest lucru ar prelungi foarte mult dezbaterea.	Romanian	langid	de
		gcld3	ro
Cu toate acestea, Václav Klaus ne cere tocmai acest lucru.	Romanian	langid	se
		gcld3	ro
Este un drept fundamental la fel de important ca libertatea de exprimare?	Romanian	langid	ro
		gcld3	ca
Prin urmare, am votat favorabil.	Romanian	langid	ro
		gcld3	it
Votul va avea loc mâine.	Romanian	langid	ro
		gcld3	et

Table 4: Results from langid and gcld3 libraries

Category	**Text**
Source	El sector agrícola tiene importantes efectos directos en la biodiversidad y los ecosistemas.
Reference	Sectorul agricol are un impact direct și semnificativ asupra biodiversității și a ecosistemelor.
Output System 1	sectorul agricol are efecte directe semnificative asupra biodiversității și ecosistemelor.
Output System 2	sectorul agricol are efecte directe semnificative asupra biodiversității și ecosistemelor.
Output System 3 (JoeyNMT)	Sectorul agricol are efecte directe asupra biodiversității și ecosistemelor.
Source	Además, la revisión general de la Directiva 2011/65/UE que debe efectuar la Comisión a más tardar el 22 de julio de 2021 debe incluir el establecimiento de un plazo realista para la adopción de una decisión sobre la solicitud de prórroga de una exención por parte de la Comisión, antes de la expiración de la correspondiente exención.
Reference	În plus, revizuirea generală a Directivei 2011/65/UE, care urmează să fie realizată de Comisie până cel târziu la 22 iulie 2021, ar trebui să includă precizarea unui termen realist pentru adoptarea de către Comisie a unei decizii cu privire la o cerere de reînnoire a unei derogări, înainte de expirarea derogării în cauză.
Output System 1	în plus, revizuirea generală a osirectivului nr.2011/65/gestionare care trebuie efectuată până cel târziu la procedura de lichidare de către lichiditatea din 1962 trebuie să includă stabilirea unui termen realist pentru adoptarea unei decizii cu privire la cererea de prelungire a scutirii de taxe vamale înainte de expirarea scutirii respective.
Output System 2	în plus, revizuirea generală a ierarhiei, „2011/65”, care trebuie efectuată cel mai târziu de islamie, la oraşul franco-demojulio de la data de 2021, trebuie să includă stabilirea unui termen realist pentru adoptarea unei decizii privind cererea de prelungire a unei scutiri de către ’omisie înainte de expirarea scutirii respective.
Output System 3 (JoeyNMT)	În plus, revizuirea generală a directivei-2011/65/UE care trebuie efectuată de Comisie până la 22 iulie cel târziu, trebuie să includă stabilirea unui termen realist pentru adoptarea unei decizii privind cererea de prelungire a exceptării de către Comisie, înainte de expirarea scutirii.

Table 5: Examples sentences from Dev set and corresponding translations. We show the difference in translation between our best system, and the system that uses a 4k vocab_size, to observe the impact of vocab_size on the translation.

Category	Text
Source	Los beneficiarios del Tratado de Marrakech son las personas ciegas, las personas que tienen una discapacidad visual que no puede corregirse para darles una función visual sustancialmente equivalente a la de una persona sin ese tipo de discapacidad, las personas que tienen una dificultad para percibir o leer, incluida la dislexia o cualquier otra dificultad de aprendizaje que les incapacita para leer obras impresas en una medida sustancialmente equivalente a la de una persona sin esa dificultad, y las personas que, por una discapacidad física, no pueden sostener o manipular un libro o centrar la vista o mover los ojos en una medida normalmente aceptable para la lectura, siempre que, como consecuencia de tales discapacidades o dificultades, dichas personas no sean capaces de leer obras impresas en una medida sustancialmente equivalente a la de una persona sin esas discapacidades o dificultades.
Reference	Beneficiarii Tratatului de la Marrakesh sunt persoane nevăzătoare, persoane care au deficiențe de vedere ce nu pot fi corectate pentru a obține o funcție vizuală sensibil echivalentă cu cea a unei persoane fără astfel de deficiențe, persoane care au un handicap de percepție ori dificultăți de citire, inclusiv dislexie sau orice altă dizabilitate de învățare, care le împiedică să citească opere imprimate în aceeași măsură, în esență, ca persoanele fără astfel de dizabilități și persoane care suferă de o dizabilitate fizică ce le împiedică să țină în mână ori să manipuleze o carte sau să își concentreze privirea ori să își miște ochii astfel încât să poată citi, în măsura în care, ca urmare a acestor deficiențe sau dizabilități, acele persoane nu pot citi opere tipărite în aceeași măsură, în esență, ca o persoană care nu este afectată de astfel de deficiențe sau dizabilități.
Output System 1	persoanele care au un handicap vizual şi care nu pot fi corectate pentru a le da o funcţie vizuală echivalentă cu cea a unei persoane fără acest tip de dizabilitate, persoanele care au o dificultate de a percepe sau de a citi, inclusiv dislexia sau orice altă dificultate de învăţare care le este imposibil de citit operele tipărite într-o măsură substanţială echivalentă cu cea a unei persoane fără această dificultate, şi persoanele care, prin handicap fizic, nu pot
Output System 2	o persoană care are un handicap vizual care nu poate fi corectată pentru a le da o funcţie vizuală substanţială echivalentă cu cea a unei persoane fără astfel de tipuri de handicap, persoanele care au dificultăţi în perceperea sau citirea, inclusiv dizolvarea sau orice altă dificultate de învăţare care le incapacită să citească opere imprimate într-o măsură
Output System 3 (JoeyNMT)	Beneficiarii Tratatului de la Marrakech sunt persoanele cu handicap, persoanele care au un handicap vizual care nu pot fi corectate pentru a le oferi un rol vizual substanţial echivalent cu cel al unei persoane fără astfel de handicap, persoanele care au o dificultate de a percepe sau lea, inclusiv dislexia sau orice altă dificultate de învăţare care le pot incapacitatea de a citi lucrări în mod fizic care, în
Source	A fin de garantizar la aplicación uniforme del presente artículo, la AEVM podrá elaborar proyectos de normas técnicas de regulación para especificar con más detalle la información que se deberá facilitar a las autoridades competentes durante la solicitud de registro tal y como se establece en el apartado 1 y para especificar con más detalle las condiciones que se establecen en el apartado 2.
Reference	Pentru a asigura aplicarea uniformă a prezentului articol, AEVMP poate elabora proiecte de standarde tehnice de reglementare pentru a preciza și mai mult informațiile ce trebuie furnizate autorităților competente în cererea de înregistrare astfel cum este prevăzută la alineatul (1) și pentru a preciza condițiile astfel cum sunt prevăzute la alineatul (2).
Output System 1	pentru a asigura aplicarea uniformă a prezentului articol, comitetul de standardizare poate elabora proiecte de norme tehnice de reglementare pentru a preciza în detaliu informaţiile care trebuie furnizate autorităţilor competente în timpul cererii de înregistrare prevăzute în alin. (1) şi pentru a preciza în detaliu condiţiile prevăzute în alin. (2).
Output System 2	în scopul asigurării aplicării uniforme a prezentului articol, μ poate elabora proiecte de standarde tehnice de reglementare pentru a preciza în detaliu informaţiile care trebuie furnizate autorităţilor competente în timpul cererii de înregistrare prevăzute la alin. (1) şi pentru a preciza mai detaliat condiţiile prevăzute la alin.
Output System 3 (JoeyNMT)	Pentru a asigura aplicarea uniformă a prezentului articol, AEVMP poate elabora proiecte de norme tehnice de reglementare pentru a specifica mai detaliat informaţiile care trebuie furnizate autorităţilor competente în timpul cererii de înregistrare în conformitate cu alin. (1) şi pentru a specifica mai detaliat condiţiile stabilite în

Table 6: Examples sentences from Test set and corresponding translation. We show the difference in translation between our best system, and the system that uses a 4k vocab_size, to observe the impact of vocab_size on the translation.

Category	Text
Source	Actualmente, se dispone de nuevas herramientas que pueden facilitar la presentación de capacidades y cualificaciones utilizando distintos formatos, digitales y en línea.
Reference	În prezent, sunt disponibile noi instrumente care pot facilita prezentarea aptitudinilor și a calificărilor folosind diverse formate online și digitale.
Our System	În prezent, există noi instrumente care pot facilita prezentarea capacităţilor şi calificărilor prin diferite formate, digitale şi online.
Google Translate	Sunt disponibile acum instrumente noi care pot facilita prezentarea competențelor și calificărilor folosind diferite formate, digitale și online.

Table 7: Examples sentence from Test set and corresponding translations. We show the difference in translation between our best system, and google translate.

	vocab_size	N	d_model	dff	h	d_k	d_v	P_drop	train steps	VAL Set		TEST Set	
										BLEU	TER	BLEU	TER
Systems A	4k	6	512	2048	8	64	64	0.1	20k	23	0.766	22.9	0.775
	8k	6	512	2048	8	64	64	0.1	15k	21.6	0.766	20.8	0.772
	16k	6	512	2048	8	64	64	0.1	20k	**24.8**	**0.693**	**25.6**	**0.689**
	32k	6	512	2048	8	64	64	0.1	20k	22.5	0.71	22.4	0.714
	48k	6	512	2048	8	64	64	0.1	15k	21.8	0.732	22.1	0.738
Systems B	4k	6	256	2048	8	64	64	0.1	20k	15.8	1.39	15.5	1.364
	8k	6	256	2048	8	64	64	0.1	20k	22.7	0.74	22.8	0.743
	16k	6	512	4096	8	64	64	0.1	25k	23.1	0.718	23.7	0.723
	16k	6	1024	4096	16	64	64	0.3	10k	22.4	0.726	22.3	0.735
	8k	6	128	2048	8	64	64	0.3	20k	19.2	0.787	18.6	0.797

Table 8: Systems A refer to the experiments that aim to show the impact of the vocab size on BLEU score
Systems B refer to the experiments that aim to show the impact of the changing parameters inside the model, with the vocab size that get the best results from Systems A (16k in our case)

	vocab_size	N	d_model	dff	h	VAL Set				TEST Set			
						BLEU	chrF	chrF++	TER	BLEU	chrF	chrF++	TER
Systems A	4k	6	256	1024	4	24.6	55.7	51.6	62.4	24.6	56.6	52.2	62.6
	8k	6	256	1024	4	25.6	57.0	52.8	61.4	25.4	58.1	53.6	61.2
	16k	6	256	1024	4	27.1	**58.5**	**54.4**	59.8	**27.5**	59.4	**55.1**	**59.5**
	24k	6	256	1024	4	**27.3**	**58.5**	**54.4**	**59.6**	27.1	59.3	55.0	59.7
	37k	6	256	1024	4	26.5	58.2	54.1	60.2	27.1	**59.5**	**55.1**	59.8
Systems B	16k	4	512	2048	8	28.6	59.5	55.5	58.5	28.9	60.7	56.4	58.1
	16k	8	512	2048	8	28.7	59.5	55.5	58.6	29.3	60.7	56.5	57.7
	16k	6	512	4096	16	28.7	59.4	55.5	58.7	29.6	60.9	56.7	57.7
	16k	6	1024	4096	8	0.3	10.6	9.0	96.4	0.2	11.4	9.8	96.4
	16k	6	512	2048	8	**29.2**	**59.8**	**55.8**	**58.3**	**29.7**	**61.3**	**57.0**	**57.4**

Table 9: Experiments in JoeyNMT
Systems A refer to the experiments that aim to show the impact of the vocab size on BLEU score
Systems B refer to the experiments that aim to show the impact of the changing parameters inside the model, with the vocab size that get the best results from Systems A (16k in our case)

	vocab_size	N	d_model	dff	h	VAL Set				TEST Set			
						BLEU	chrF	chrF++	TER	BLEU	chrF	chrF++	TER
System A	4k	6	256	1024	8	11.5	36.7	33.3	81.9	13.5	40.5	36.9	77.6
System B	4k	6	256	1024	8	24.6	55.7	51.6	62.4	**24.6**	**56.6**	**52.2**	**62.6**

Table 10: System A refers to the system that uses all data. System B refers to the one that uses data filtered by langid library.

5. Training ideas for relevant experiments with Transformers to the proposed problem

6. Establishing some points to define the cleaning of the data collection, its analysis and the pre-processing step that would help the training stage

Gîdea Andrei

1. Researching and writing code for the RNN architecture with attention layers for the end-to-end translation system

2. Training ideas for relevant experiments with RNN to the proposed problem

3. Training experiments with NMT Joey and Hugging Face frameworks for comparisons our implementation

4. Writing the documentation and the presentation

5. Establishing some points to define the cleaning of the data collection, its analysis and the pre-processing step that would help the training stage

8 Conclusions and Future Work

Working on this project, we better understood the concepts behind the transformer-type architecture, which we might not have understood if we were just reading some articles, without actually going into the code part.

In addition to the parallel data used in this paper, a major improvement may be the use of monolingual data, which can be used for the train of embedding to better learn the words in the vocabulary used, thus helping the final system.

Also, one aspect that could bring better results would be to bring the data into truecase, or use it as such, instead of making it lowercase(in the case of our framework).

In addition, we could experiment with other methods of data tokenization, such as sentencepiece (Kudo and Richardson, 2018) or why not, defining our own tokenization rules, which would better fit the data set used.

References

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the unseen? yorùbá → english MT in low-resource, morphologically-unmarked settings. *CoRR*, abs/2103.04225.

Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. 2020. Translating similar languages: Role of mutual intelligibility in multilingual transformers. In *Proceedings of the Fifth Conference on Machine Translation*, pages 381–386, Online. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *CoRR*, abs/2004.14928.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 36–46, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Europarl: A parallel corpus for statistical machine translation. 5.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 109–114, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages

25–30, Jeju Island, Korea. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, and Dan Tufis. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.