Idiomatic Language Translation and Transfer Learning

Stanford CS224N Custom Project
Mentor: Hugh Zhang
Final Project Option 3

Jupinder Parmar
Department of Mathematics
Stanford University
jsparmar@stanford.edu

David Estrada-Arias
Department of Computer Science
Stanford University
dae783@stanford.edu

Abstract

Current translation models fall short for a number of translation patterns. One such example is idioms, and in this project we hope to provide better translation of idiomatic expressions within languages by improving upon an implementation of a Neural Machine Translation (NMT) model. We aim to do so by altering the attention mechanism and testing different loss functions in our architecture. Additionally, since state of the art machine translation models take long amounts of time to train, we explore transfer learning as a method to reduce total training time between languages. Specifically, we first train a NMT between English and the Latin script Transliteration of Greek and use this NMT to initialize our parameters on a translation model between English and Greek. We find that the Dual Skew Divergence Loss allows helps in improving the translation accuracy of idiomatic expressions while our transfer learning architecture lead to similar BLEU scores yet took about half the amount of training time compared to a regular English-Greek NMT model.

1 Introduction

When learning a new language, recognizing idiomatic expressions is often the toughest task on the way to mastering the language. We see that it often takes many years of familiarity with a language before an individual is able to understand such expressions. Because of this, phrases lost in translation such as idioms present interesting problems that most NMT models currently fail to translate properly.

Therefore, we think that developing a neural architecture that is able to not only understand the basic semantic underpinnings of source sentences but also of these idiomatic expressions will be able to allow for better translations overall. In the field of machine translation, there has not been high focus on idiomatic expressions simply because of their inherent complexity. We see that limited literature has been published on this area and the papers that are out there focus on building data sets for this task while applying vanilla NMT models [5]. The success that these papers have seen is highly dependent upon their carefully curated data sets consisting of idioms as they train their model on idioms explicitly. However, the inherent issue with building such complex data sets is that it often relies upon having a professional translator and takes a while to generate enough parallel sentences to effectively train a NMT on.

Thus, within our approach to improving the translation of idiomatic expressions between languages we seek to find improvements on the NMT model rather than creating a data set for a specific language. Specifically, we will be focusing on translating phrases between Spanish and English. In our attempts to create an NMT model that meets these specifications we focus on applying a different attention mechanism that is able to encapture more complexity in its operations, namely additive attention[1]. We also implement a different loss function than the current cross entropy loss which is standard in

most NMT systems, namely the Dual Skew Divergence Loss that has been shown to provide better performance and generalizability for NMT systems [9]. Lastly, we also work towards implementing a contextual pre-trained word embedding model such as ELMO[8] because later within our paper we will show that context is an important factor within idiom translation. From our results of these tests we notice that altering the attention mechanism doesn't have much of an impact on idiomatic translation, but switching to the dual skew divergence loss increases BLEU scores by almost a whole point on idiomatic expressions.

In addition to translating idioms, we noticed that as we were training our NMT models they often took quite a while to train thus we seek to analyze transfer learning as a method of reducing training time of neural machine translation systems. Transfer learning has been investigated in depth within machine translation systems but is more often analyzed as a method with which to create models for low resource languages[10]. And, as literature within this field has shown, transfer learning has been quite successful in improving the accuracy of translations with languages that contain sparse parallel corpora. Thus, we believe that applying transfer learning can help reduce the total train time of our NMT architecture. Specifically, we seek to train a neural architecture on the task of translating sentences between English and the Latin script transliteration of Greek. We hope to utilize this model based on the English transliteration of Greek to initialize our parameters for a regular Greek to English NMT model.

Implementing this transfer learning based approach using a small parallel corpora of English and the Latin script transliteration of Greek we were able to obtain similar results in terms of BLEU scores when initializing our English-Greek NMT model with the parameters of transliteration based NMT in comparison to those of a regular English-Greek NMT Model. However, we see that by using our transfer learning architecture we took only 10755 seconds in training time while the regular NMT model took 24876 seconds in training time – a reduction in training time by more than half.

2 Related Work

There have been a couple recent pieces of literature that are related to the domain that we are working in. Specifically, in this section we will give a brief overview of three papers and discuss their relation to the tasks that we are trying to accomplish.

2.1 Generative Neural Machine Translation

The Generative Neural Machine Translation(GNMT) [2] model is a related machine translation model to the NMT. In particular the GNMT model differs from a traditional NMT in that it hopes to model the semantics of the source and target sentences by introducing a latent variable. This latent variable is introduced as a means to create a better semantic representation of the sentence allowing for more accurate translations.

The GNMT accomplishes this by modeling the joint distribution of the target and source sentences instead of the conditional distribution of the target sentence given the source sentence which most NMT models follow. It is able to do so by conditioning upon this latent variable which acts as a language agnostic representation of the sentence in both languages. By giving the latent representation responsibility for generating the same sentence in multiple languages, it focuses on learning the semantic meaning of the sentence resulting in better understanding of the sentence.

We see that the goal of the GNMT to create a better understanding of the sentence it is translating in order to create translations with higher semantic similarity is highly related to our motivation behind seeking to accurately translate idioms. We believe that effective translation of idioms within a language will give us better translation of sentences in that language overall as we are able to understand the most complex phrases that exist in the language. We seek to compare the results of our methodologies with those of the GNMT in order to potentially determine if the GNMT would be a viable avenue to consider for translation of idioms.

2.2 Current Models in Idiom Translation

As mentioned in the introduction, most literature that exists for idiom translation focuses on building custom data sets with which to train vanilla NMT models on to produce results for this task. The

leading paper that exists in this area, [5], focuses on creating a data set of German-English idiomatic sentences. They end up developing a parallel corpora of over 4.5 million sentences that contains over 2k sentences with idiomatic expressions and over 130 unique idiomatic phrases. Through the development of this corpora and training a regular German-English NMT on this curated datset they were able to achieve relatively high BLEU scores for such a complicated task.

While working in the same area as this paper, we will approach our problem in a different manner as we would like our model architecture to be generalizable and successfully translate idioms contained within a variety of languages. Thus, we see that we are essentially taking the motivation behind the GNMT and essentially trying to formulate a similar model for idiomatic translation, both orthogonal to the motivation behind [5]. Thus, we focus on finding the best NMT architecture for idiomatic translation instead of curating a specialized data set for training our model. We believe that by following this approach we also will be able to determine what NMT architectures will lead to higher semantic meaning within translations.

2.3 Transfer Learning for Low Resource Languages

We see that Zoph, eta al[10] use transfer learning to combat the issue that has been shown with the traditional encoder-decoder framework for NMT of not being effective for low-resource languages. The method that they employ within their architecture is to train a high-resource language pair then transfer some of the learned parameters to the low-resource pair to initialize training. By doing so they were able to improve BLEU scores of baseline NMT models by a significant margin on low-resource language pairs.

As we seek to reduce the amount of training time of our NMT models, we will borrow the concept of initializing our final model with the pre-trained parameters of another model that was exemplified in this paper. However, instead of our initial model being between a high resource language, we will train it on the Latin script transliteration of Greek. Motivated by the success of [10] we believe that by doing so we can effectively lead the parameters of the Greek-English NMT closer to their final optimal values and decrease training time overall.

3 Approach

3.1 Baselines

As no piece of literature has taken our approach to the problem of idiom translation we develop our own baseline. We utilize and build upon the vanilla encoder-decoder NMT system we developed in assignment 4 which was created by the CS224N TAs. We train our Spanish-English NMT system on data obtained from [4] and test it on our own idiom data sets that we created.

For our transfer learning baseline, we simply train the same vanilla NMT architecture on a English-Greek corpus collected from [4] to determine BLEU scores as well as how long this model takes to train. We also investigate the performance of this architecture on a related low resource language, Ancient Greek, whose sparse corpus we again were able to obtain from [4]. The results of our baseline models are shown below in Table 1.

	BLEU score	Perplexity
Greek-English test set (10% of Wikipedia corpus)	28.5027	20.463
Regular NMT on Ancient Greek Data	7.255	271.89
Spanish-English independent idioms (i.e. w/o context)	2.121	3365.14
Spanish-English sentences w/ idioms (i.e. w/ context)	11.2703	55.572

Table 1: **Baseline.** The results from the four tests we completed after training our models with the data sets obtained from [4].

From these results we are able to gather that our baseline NMT system was able to perform better on idiom translation when it had context surrounding the phrase rather than simply translating the phrase alone as there is a high disparity in the BLEU scores between idiomatic sentences vs phrases. Additionally, we note that idiomatic sentences have much lower BLEU scores than regular sentences in NMT systems. Seeing that context is a helpful feature to include for our problem, we will focus in

implementing and trying out approaches that will take advantage of the context that words are present in. Additionally, for our Greek-English NMT system we note that it takes 24876 seconds of total training time while our Ancient Greek-English NMT system does not perform particularly well due to the limited training data that was available.

3.2 Additive Attention

The first methodology that we will try out to improve the performance of our baseline NMT system on translating idiomatic expressions will be to switch from multiplicative attention that is currently present in the architecture to additive attention. Attention in general is a mechanism to obtain a fixed-size representation of an arbitrary set of representations which has been used to increase the performance of a variety of deep learning architectures. Essentially, within an encoder-decoder system, we see that using attention allows for us to determine which words in the source sentence are most related to the current word we are decoding.

As discussed in lecture and [1], we see that additive attention is a more complicated form of attention than multiplicative attention. The attention scores for both types are computed as follows respectively :

$$\mathbf{e}_i = \mathbf{b}^{\mathrm{T}}\tanh(\mathbf{W}_1\mathbf{h}_i + \mathbf{W}_2\mathbf{s}) \tag{1}$$

$$\mathbf{e}_i = \mathbf{s}^{\mathrm{T}}\mathbf{W}\mathbf{h}_i \tag{2}$$

In (1), additive attention, we get that $\mathbf{W}_1$, $\mathbf{W}_2$ are weight matrices while $\mathbf{v}$ is a weight vector. In comparison we see that in (2), multiplicative attention that only $\mathbf{W}$ is a weight matrix. By including not only more operations and parameters to learn through the additional weight matrix and weight vector, but also the non linearity through the *tanh* function we see that additive attention is able to model a more complex relationships between our hidden encoder and hidden decoder states.

We will implement this formula for additive attention within our NMT system to determine how it will help with the translation of idiomatic expressions. We believe that this will be an effective tool in helping with this task because the additional matrix and vector may be able to model a deeper relationship between a phrase and the context that surrounds it which would be helpful knowledge for our NMT system in translating idioms as shown in our baseline results.

3.3 Dual Skew Divergence Loss

Additionally, we seek to improve upon our baseline NMT by changing the loss function used within training. Implementing a loss function that penalizes incorrect translations of idiomatic expressions higher than those of regular expressions should lead to a model that can translate idioms effectively. Creating such a loss function that recognizes idiomatic expressions would be very difficult, so we turn to [9] and implement the Dual Skew Divergence Loss which is derived as shown below.

We first begin with KL divergence which is given by $D_{KL}(P||Q) = E_{x\sim P}[\log P(x) - \log Q(x)]$. From this we are able to obtain Skew Divergence which takes a parameter $alpha$ and is given by $s_\alpha(P, Q) = D_{KL}(Q||\alpha Q + (1-\alpha)P)$. Through this we obtain the final formula for Dual Skew Divergence given by $D_{DS} = \beta s_\alpha(P, Q) + (1-\beta)S_\alpha(Q, P)$ which depends upon a parameter β. The explicit formula for the Dual Skew Divergence loss which we implement in our architecture is shown below:

$$J_{DS} = -\frac{1}{n}\sum_{i=1}^{n}[\beta\mathbf{y}_i\log((1-\alpha)\hat{\mathbf{y}}_i + \alpha\mathbf{y}_i) - (1-\beta)\hat{\mathbf{y}}_i\log(\hat{\mathbf{y}}_i) + (1-\beta)\hat{\mathbf{y}}_i\log((1-\alpha)\mathbf{y}_i + \alpha\hat{\mathbf{y}}_i)]$$

As shown in [9] we see that building upon KL divergence which measures the difference between two probability distributions, dual skew divergence loss is able to give a better tradeoff between generalization ability and error avoidance during NMT training. Utilizing this loss within NMT systems has lead to higher BLEU scores than other traditional loss metrics.

This generalization ability compared to the regular cross entropy loss implemented in most NMT systems is especially useful for our task where we are seeking to correctly translate phrases that aren't the most commonly used. We believe that this loss will be able to cater to our needs of emphasizing semantic understanding of a sentence which is especially useful within idiomatic translation. Within our implementation we use $\alpha = .01$ and $\beta = .10$ as suggested by [9] for optimal performance..

3.4 Transfer Learning

We have briefly highlighted our methodology that we are taking for our transfer learning approach above, but will go into more explicit detail here.

We first will create a data set of the Latin script transliteration of a subsection of the Greek Corpus that we trained our baseline model on above. We go into more detail about this in our Data section below. Then, we will take our baseline NMT architecture consisting of a encoder-decoder LSTM model and train it on the parallel corpora of Latin script transliteration of Greek and their corresponding English sentences. We hope by doing so we will learn a better representation of our initial parameters of a Greek-English NMT architecture than randomly initializing the system. In addition as this transliteration data set will be smaller than the original Greek one the training time for this model should be less.

After training this initial NMT architecture, we will take its weights and initialize our NMT architecture for Greek-English translation with them. We then will train this NMT model with the weights given by our transfer learning approach on the original Greek corpora. We believe that the Latin script transliteration of a language should give important information to a machine translation model about the relationship between the two languages that can be then further built upon which is how we came up with our idea. This transfer learning approach on training an initial model on the transliteration of a language is completely original.

4 Experiments

4.1 Data

For the baseline Spanish-English seq2seq model, we downloaded a Wikipedia corpus from Opus [4] that contains over 1.8 million Spanish-English pairs of sentences. Due to continually memory errors on our VM when training on this large corpus, we decided to cut this data set down. We use a training set of 150 thousand sentence pairs, a validation set of one thousand sentences, and a test set of 10 thousand sentences.

Additionally, we obtained a Wikipedia corpus [4] from Opus that contained over 180 thousand pairs of Greek and English sentences. We utilized this dataset to train our baseline Greek-English seq2seq model as well as our final Greek-English NMT model within our transfer learning approach. Just like we did for our Spanish corpus, we split the Greek corpus into a train set of 150 thousand sentence pairs, a validation set of one thousand sentences, and a test set of 10 thousand sentences. Additionally, as we seek to determine the effect of our transfer learning approach on low resource languages we obtain a Tatoeba corpus from [4] on Ancient Greek-English that consists of 600 total sentence pairs. We split this into a train set of 450 sentence pairs, a validation set of 50 sentence pairs, and a test set of 100 sentence pairs.

As there was no available Spanish-English idiom data set to test our model, we hand-made our own Spanish Idioms data set with around 360 common Spanish Idioms and their English translation using an online database [2]. Additionally, we hand-wrote 100 idiomatic Spanish sentences along with their translation as another data set. Having both data sets will allow us to determine how context impacts the performance of idiom translation. Here are some examples of Spanish idioms and their translations:

Spanish phrase	Literal English Translation	Correct English Translation
ojo al parche	look out	eyes on the patch
al hambre no hay pan duro	hunger does not have hard bread	beggars can't be choosers
a que santo encomendarse	to which saint to entrust oneself	at a loss

Some examples of our hand made idiomatic Spanish sentences are shown below where the idiom is in bold within the sentence:

Spanish Sentence	Correct English Translation
ella anda a **paso de tortuga** porque es muy vieja.	she tends **to move slowly** because she is so old.
haz una **vuelta a la manzana** y despues ven a casa.	go **around the block** and then come home.

As no corpus of the Latin script transliteration of Greek is available, we additionally made our own transliteration data set. We took 22.5k pairs of sentences from our original Greek corpora then used [6] to find the Latin script transliteration of all of the Greek sentences. After doing so we split our corpora into a train set of 20 thousand sentences pairs, a validation set of one thousand sentence pairs, and a test set of 1.5 thousand sentence pairs. Some example sentences from this data set are shown below:

Transliteration of Greek	English Sentence
Afierose epises kapoio chrono sten politike ste Vrazilia.	He also spent some time in politics in Brazil.
E parayoye kermaton stamatese to 1862..	Coinage production was suspended in 1862.

4.2 Evaluation Methods

To evaluate our translation, we utilized BLEU and perplexity scores, which were introduced to us earlier in CS224N coursework. Perplexity scores allow us to see how well our model is able to recognize patterns in characters/words. The BLEU method [7] compares the machine translation with the provided reference translation(s), focusing on the n-gram precision. We utilize these metrics since they have become the (nearly) universal way to measure the success of a NMT model.

4.3 Experimental Details

Our baseline experiments consisted of training a Spanish-English NMT on the relevant data, as well as a Greek-English NMT on the data pertaining to it. Both of our NMT architectures consist of an LSTM sequence to sequence encoder-decoder that utilize a dropout layer. In both models we used a batch size of 2 on the train set and a batch size of 4 on the dev set and begin with a learning rate of .001 utilizing a learning rate decay of 0.5. We see that our initial Spanish NMT took a total trainign time of 10.78 hours while our Greek NMT took 6.9 hours to train.

We then created models in alignment with the approaches discussed above for idiom translation. We were able to fully implement a model that used additive attention and a model that made us of the dual skew divergence loss. We initialized both of these models with the same parameters and batch sizes of our baseline model to isolate the true effectiveness of each of these proposed changes. We see that the NMT with additive attention took 9.56 hours to train while the NMT model with dual skew divergence loss took 9.78 hours to train.

For our additional models on the transfer learning side of our project we first trained an NMT model on our transliteration data set that we created. We kept the same initialization for parameters as the vanilla NMT model for our baseline which we mentioned above. We see that this model took 1.7 hours to train or 6087 seconds. We then took the parameters of this transliteration model and used them to initialize our weights of our Greek-English NMT model. All other parameters such as learning rate and batch size remained the same as in our baseline models. Training on our Greek data set we see that this model took 4668 seconds or 1.3 hours. Additionally, as we thought it would be interesting to see if our transliteration transfer learning approach could act as a methodology to translate low resource languages we initialized an Ancient Greek-English NMT system with our transliteration NMT parameters and trained this on our Ancient Greek Corpus. as Our Ancient Greek corpus was small it took a total training time of .77 hours or 2783 seconds.

We note that all of our models were trained on the Microsoft Azure NV6 VM.

4.4 Results

After training our models, we tested them on their related testing set and recorded their BLEU and perplexity scores. Review Table 2 and Table 3 for a complete breakdown of our subsequent results in our Spanish and Greek data sets.

	BLEU score	Perplexity
Additive Attention for independent idioms (i.e. w/o context)	2.312	2699.66
Additive Attention for sentences w/ idioms (i.e. w/ context)	10.87	78.76
Dual Skew Divergence Loss for independent idioms (i.e. w/o context)	3.512	1899.43
Dual Skew Divergence Loss for sentences w/ idioms (i.e. w/ context)	11.87	38.76

Table 2: **Spanish.** The results from the tests we completed for our Spanish data sets.

	BLEU score	Perplexity
Transliteration Model(Latin script transliteration of Greek)	22.8943	25.2456
Greek-English NMT Initialized with Transliteration Model	29.3031	21.317
Ancient Greek-English NMT Initialized with Transliteration Model	6.698	928.928

Table 3: **Greek.** The results from the tests we completed for our Greek data sets.

We see that in both of our implemented approaches, additive attention and dual skew divergence loss, that we still see a large disparity in BLEU and perplexity scores between idiomatic sentences vs idiomatic phrases demonstrating the usefulness of context yet again. Analyzing these results we see that additive attention generally had the same BLEU and perplexity scores as the baseline while our dual skew divergence loss model was able to improve upon our baseline model in all aspects. This demonstrates that the generalization ability of the dual skew divergence loss was a helpful characteristic to embed within an NMT model for idiomatic translation. We believe this occurs since idioms are generally not often represented much within text that the higher semantic understanding given by this new loss function allows for better idiomatic translations. We note that while we were expecting the additive attention model to make better use of context than multiplicative attention that this was not the case in the results of our model. We suspect that this is because learning to make better use of context in idiomatic sentences might be too complicated of a task even for the more complex additive attention.

For our results on the transfer learning side of this project, we first recognize that our transfer learning, transliteration based approach for Greek-English translation took a total of 10755 seconds or 2.98 hours to train. We see that this a large reduction in training time from the 6.9 hours it took the regular Greek-English NMT model to train. Additionally, we note that our transfer learning architecture was able to achieve a higher BLEU score than the vanilla architecture as well. Both of these improvements are based off the fact that the transliteration model was able to create a much better start for our model parameters of the Greek-English NMT model than a random initialization. Interestingly however, this transfer learning based approach for Ancient Greek did not lead to improvements, but rather a decline in BLEU scores. We suspect that this happened because the transliteration model was based off of Greek and not Ancient Greek meaning that it does not perform well with related languages, only the specific language we created the transliteration for itself.

5 Analysis

Below we see the results of various English translations of Spanish idiomatic sentences from all of the architectures that we created in this project as well as the gold standard translation.

Regular Spanish-English NMT: (1) I don't care whether I win or lost. (2) Get back to the apple and then you see home.

Additive Attention Spanish-English NMT: (1) I don't care whether I win or lost at end. (2) Get back to the apple and then you see home.

Dual Skew Divergence Loss Spanish-English NMT: (1) At the end, I don't care whether I win or lost. (2) Get back to the block and then you see home.

Gold Standard English Translation: (1) At the end of the day, I do not care if I win or lose. (2) Go around the block and then come home.

These translation demonstrate how difficult it can be to successfully translate idioms. Often times a model's ability to do so is hit or miss, as evidenced by translations (1) and (2) across the three NMT models. The most successful translations were completed by the Dual Skew Divergence Loss NMT, where it at least understood the context of the sentences and made more or less successful translations. For example, for sentence (2), this NMT understood the main idea of "the block" instead of literally translating the word "manzana", or "apple". Also, while the other two models have a similar translation for sentence (1), the Dual Skew Divergence Loss NMT produced the translation closest to the Gold Standard, as well as the most coherent one of the three. Clearly this model was able to understand the meaning of the idioms better than the others, since they still struggle with literal translations and do not capture overall meanings very well.

Below we see the results of various English translations of Greek sentences from all of the architectures that we created in this project as well as the gold standard translation.

Regular Greek-English NMT: (1) *The version of Jackson was released as a single in the US to promote "The Original Soul of Michael Jackson".* (2) *The main number of products are wine and the band.*

Transfer Learning Based Greek-English NMT: (1) *The Jackson version was released as a single in the US to promote "The Original Soul of Michael Jackson".* (2) *The main farming products are wine and cheese.*

Gold Standard Translation: (1) *Jackson's version of "Twenty-Five Miles" was released as a single in the US to promote "The Original Soul of Michael Jackson".* (2) *The main agricultural products are wine and cheese.*

Through these translations we see that both of our architectures performed similarly on sentence (1) while missing out on a key phrase that was include in the gold translation; however on sentence (2) we see that our transfer learning architecture translation almost completely matches the gold standard while the regular NMT has a number of inaccuracies in its translation. We see that these output translations show that our transfer learning architecture has the ability to outperform a regular Greek-NMT system on perhaps more complicated words because it is able to correctly understand the meaning of certain words such as *agriculture* which it choose to translate with *farming* that the regular NMT couldn't understand. However, we see that both models we created did not include *"Twenty Five Miles"* in their translation of sentence (1) showing that our transfer learning model still struggles with proper nouns.

6 Conclusion and Future work

Within our project we were able to recognize that context is a key factor in helping translate idiomatic expressions. Additionally, we note that an NMT with dual skew divergence loss is the current best improvement we have on an NMT system to produce better BLEU and perplexity scores on idiomatic sentences/phrases. Lastly, we have developed an efficient transfer learning based architecture where we train on the Latin script transliteration of a language before initializing our final NMT with those parameters to help reduce the total amount of training time while slightly increasing BLEU scores.

We tried to develop an architecture that made use of ELMO embeddings, however we were unable to get it successfully working in the limited time frame that we had. We believe that such a model would be helpful for idiomatic translation and is an area for future work(we include our failed code as a starting point). Additionally, as the GNMT system seeks to represent the semantic underpinnings of a sentence we believe that it would be a useful architecture to apply towards idiomatic translation in the future as well.

References

[1] Galassi, et al. Attention, please! A Critical Review of Neural Attention Models in Natural Language Processing. *In Arxiv*, 2019.

[2] Hashil Shah and David Barber. Generative neural machine translation *In Neural Information Processing Systems (NeurIPS)*, 2018.

[3] Hugh Fox, 244 Spanish Idioms, *https://foxhugh.com/spanish/244-spanish-idioms*, 2014.

[4] Krzysztof Wołk and Krzysztof Marasek: Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs., Procedia Technology, 18, Elsevier, p.126-132, 2014.

[5] Marzieh Fadaee and Arianna Bisazza and Christof Monz. Examining the Tip of the Iceberg: A Data Set for Idiom Translation. *In Arxiv*, 2018.

[6] My Languages.org, *http://mylanguages.org/greek_romanization.php* , 2019.

[7] Papineni, et al. BLEU: a Method for Automatic Evaluation of Machine Translation. 2002. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318*, 2002.

[8] Peters, et al. Deep contextualized word representations. *In North American Chapter of the Association for Computational Linguistics(NAACL)*, 2018.

[9] Xiao, et al. Dual Skew Divergence Loss for Neural Machine Translation. *In Arxiv*, 2019

[10] Zoph, et al. Transfer Learning for Low-Resource Neural Machine Translation. *In Conference on Empirical Methods in Natural Language Processing*, 2016.