Performance Evaluation of Recurrent Neural Network on Large-Scale Translated Dataset for Question Generation in NLP for Educational Purposes

Fidel I. Mamani Maquera, B.Sc[1], Alfredo Paz Valderrama, Ing.[1], and Eveling G. Castro Gutierrez, M.Sc[1]

[1]*Universidad Nacional de San Agustín de Arequipa, Perú, fmamanimaq@unsa.edu.pe, apazv@unsa.edu.pe, ecastro@unsa.edu.pe*

Abstract*–In recent years, neural networks have been used widely to solve many NLP tasks that involve large-scale datasets. Recently, Question Generation (QG) has called great attention since it is a subtask of Question Answering (QA) that has many applications in the real world, mainly for educational purposes. The importance of it could be seen on many recently released large-scale datasets prepared exclusively for this task, most the data used in NLP are available in the English language, but it is not the case for the rest of the languages, like Spanish, which is the third most used language in the world. This research is focused on analyzing the performance of current state-of-the-art neural network models used in QG using translated Spanish large-scale dataset from English. To know the accuracy of the translated Spanish data from English, it has been used state-of-the-art OpenNMT machine translator and Google Translation API, then the results have been analyzed with the corresponding automatic metrics - BLEU, METEOR, ROUGE - and human evaluations such as fluency and adequacy, later, it has been trained a state-of-the-art question generation (QG) neural network model using Spanish translated data to generate automatic questions in Spanish language. Surprisingly, the results outperform the original English results in average 37% on all automatic evaluation metrics. To the best of our knowledge, this work is the first one using large-scale Spanish translated data for QG task using recurrent neural networks for educational purposes.

***Keywords*-- google translation, recurrent neural network (RNN), natural language processing, translated data, squad dataset**

I. Introduction

Recent years, the development of neural network has allowed researchers to overcome several tasks in Natural Language Processing (NLP) [1]-[7]. Such developments and interest have been increased because of the large amount of large-scale dataset prepared that exist for each task in NLP.

Most recently, Question Generation (QG), which is a task related to Question Answering (QA) has been calling great attention because there are many applications that required this problem to be solved as suggested by [1], [4], [6], [8]. In [9] this task could be helpful in hospitals to diagnose patients. In [10], [11], it is used to generate questions from images, and [4], [8], [12] suggest that one key application could be used to help students with reading comprehension materials as it was used previously by [8] in educational environments.

Question Generation (QG) is aimed to create natural interrogative sentences from given paragraphs or sentences. This goal has been improved recently by using Recurrent Neural Network (RNN) [4]-[7], and such improvements were possible because of the large amount of large-scale datasets that exist. Among all existing datasets, the most popular ones are: SQuAD [13], MS Marco [14], bAbI [15], QuAC [16], CoQA [17] and WikiQA [18]. All of them were prepared for QA task, but they could be adapted for QG task. Moreover, there are recently released datasets that were prepared exclusively for QG task such as LearningQ [12] and the adapted SQuAD dataset for QG used in [4], both [4] and [12] were prepared for educational purposes.

In contrast to other languages like Spanish, even though being the third most used language in the world. To the best of our knowledge, there is no free large-scale dataset available for QG task, there are some datasets for QA task in Spanish language, but they are small to be applied to neural network models. However, QG task have been studied for Spanish language, but they have not used neural networks [19].

As said before, the importance of large-scale data is crucial when using neural networks since it helps to get better performance as described in [12], [13]. Because of the lack of Spanish large-scale datasets for QG task, this paper proposes to translate current English large-scale dataset used for QG task using Google Translation API and OpenNMT Machine Translation toolkit [20] to translate the adapted SQuAD for QG dataset provided by [4].

This approach has been executed as follows. First, the adapted SQuAD for QG dataset provided by [4] has been translated to Spanish language. Second, the accuracy of the results have been evaluated using automatic evaluation metrics, BLEU, METEOR, and ROUGE. It has been also evaluated by human evaluation using fluency and adequacy as suggested by [8], [21]. Third, It has been used a corpus in Spanish to get vectors of words using word embedding techniques such as Vect2Word [22] and Glove [23]. Fourth, It has been trained a state-of-the-art sequence-to-sequence model to generate questions from paragraphs and sentences [4] for Spanish language. Finally, The accuracy of the results has been measured and in contrast to previous work, it has been found that using translated data with current state-of-the-art Question Generation neural network model, the results outperform the original English dataset results.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.178
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

The remainder of this paper is organized as follows. Section 2 makes enfasis on the related work. Section 3 describes all datasets used throughout this research such as European Parliament (europarlv7), Spanish Billion Words Corpus and Embeddings (SBWCE), and SQuAD Dataset. Section 4 describes the use of Google Translation API and OpenNMT to translate data from English to Spanish. Section 5 presents and evaluates the results of the translation. Section 6 describes the state-of-the-art method used to train and generate questions from sentences and paragraphs. Section 7 presents and discusses the results obtained when generating automatic questions from translated Spanish sentences and paragraphs. Finally, section 8 discuss conclusions and future work for this research.

II. Related Work

This section is divided into three parts. First, it presents relevant work done in English for QG task. Second, it describes the datasets used in Spanish NLP tasks and the research done for QG task. Finally, it presents some research that used translated Spanish dataset to accomplish their goals.

A. Question Generation(QG) Task

Many research consider two important periods on QG Task. At the beginning, computer scientists and learning scientists typically have tackled QG task by using rule-based systems as described in [4], [8], [12], [24], where the rules were defined carefully by experts, then the input sentence was transformed using these rules to be converted to an interrogative sentence. the success of this approach is dependant on the number of rules, which makes it harder to come up with all the rules that a sentence might have and the number of variations that a sentence could be written.

Recently a data-driven approach has emerged as one of the most promising techniques to tackle this problem by using recurrent neural networks, such techniques are described in [4], [5], [6], [7], but this methods are largely dependant on large-scale dataset and also the quality as it is remarked in [12], [13].

All of these research have been done using English language and furthermore the data used is also in English.

B. NLP in Spanish language

The Majority of the tasks in NLP uses small datasets. Among the most popular organizations that provide their datasets with some NLP tasks are: Cross-Language Information Retrieval (CLIR) [25], [26], Cross-Language Evaluation Forum (CLEF) [27], Sentiment Analysis at SEPLN (TASS) [28] and SEPLN [29]. Most of them focused on cross-language relation between European languages and there are some small Spanish-English data, which are released with NLP tasks for the community. The majority of them are related to Sentiment Analysis as referred in [30], [31]. These works have competed in semeval 2017 task-2 [32], which is a large-scale data for sentiment analysis from twitter. Others big data involves Parallel corpus for statistical machine translation for European languages [33].

In addition, QG task in Spanish language is still being developed by using ruled-based approach using hand crafted rules made by linguistic experts as in recent publications [19].

C. Use of translated data for NLP tasks

As it is has been described, to the best of our knowledge there is no available free large-scale dataset for QG in Spanish to apply state-of-the-art recurrent neural network (RNN) models, but there are research that used translated data to accomplish their goals. As in [30], they used Google Translation API to measure the performance of the translated data using three languages. [31] refers that had some issues with data since the benchmark for semantic similarity task in NLP is done using English dataset, so they formed their own dataset to overcome the problem. in [34], In order to measure the similarity of word embeddings they used Google Translation to match English data with Spanish.

III. Datasets used Throughout this Paper

In order to generate automatic questions from paragraphs and sentences in Spanish language using neural networks, it has been used three large scale datasets. First, adapted SQuAD for QG provided by [4]. Second, the European Parliament (europarlv7) [33], which offers many parallel corpus for Machine Translation. For the purposes of this paper, Spanish-English corpus has been used from the entire dataset. Finally, to get the word embedding vectors which uses a corpus, it has been used Spanish Billion Words Corpus and Embeddings (SBWCE) from [35].

A. Adapted SQuAD for Question Generation (QG) Task

Large-scale datasets for Question Generation (QG) help to develop new methods and techniques as stated in [12], [13]. Since there is no available large-scale dataset for Question Generation in Spanish, this research has used the adapted SQuAD for Question Generation task, this dataset has been prepared by [4], and it could be downloaded from https://github.com/xinyadu/nqg, which is in English language and comes originally from SQuAD dataset [13].

SQuAD dataset has been elaborated for Question Answering (QA) task and the paragraphs and sentences have been manually crafted from wikipedia articles, the questions where elaborated by crowdworkers, as described in [13]. (Du et. al) [4] used these paragraphs and sentences from SQuAD to pair each sentence with a question, and in this way they got an adapted SQuAD dataset for QG task to generate automatic questions.

The resulting dataset created by (Du et. al) [4] contains 92 931 paragraphs, 92 931 sentences, and 92 931 interrogative sentences. The dataset has been divided into two categories, sources and target, where each source sentence has been paired with a target question that relates to the original paragraph or sentence. Besides, as it is usual in neural network, the data has been divided into three sets: development, training, and test, as shown in Table I.

TABLE I
PROPERTIES OF THE DATASET PREPARED BY DU FOR QUESTION GENERATION TASK

Dataset	Type of Text	Number of samples(source)	Number of samples (target)
Development	Paragraph	10 570	0
Development	Sentence	10 570	10 570
Train	Paragraph	70 484	0
Train	Sentence	70 484	70 484
Test	Paragraph	11 877	0
Test	Sentence	11 877	11 877

B. *European Parliament Parallel Corpus*

Among all parallel corpus for English-Spanish that exist, European Parliament (europarlv7) [33] is the largest one, with over 1,9 million paired sentences. This corpus has been used to train a model to translate the Adapted SQuAD for QG dataset. In this way, it allowed to measure and compare the accuracy of the translation from English to Spanish.

C. *Spanish Billion Words Corpus and Embeddings*

The majority of the models that use neural network use the two most famous word embedding techniques called vect2Word [22] and Glove [23], such techniques allow to represent the words as vectors and give them a numerical value as described in [22], [23]. Such techniques has been largely used in many languages.

In order to get a better representation of words by word embedding, it is necessary a large scale dataset, commonly known as a corpus, which contains almost all the words that exist in a language. [35] has collected Spanish data from different sources to get a large Spanish corpus, this corpus contains around 1.5 billion words in Spanish, such corpus could be found at https://crscardellino.github.io/SBWCE/.

IV. TRANSLATION OF ADAPTED SQUAD DATASET FOR QUESTION GENERATION (QG) TASK

This section describes in detail all the steps used to translate adapted SQuAD dataset for QG using Google Translation API and OpenNMT Machine Translation [20].

A. *Google Translation API*

Google Translation API is one the most famous language translator known, many research have used it to measure the accuracy of their translations as in [36]-[40]. All of these research agreed that Google Translation API es better than any other language translator. Furthermore, Google Translation API is a good choice in order to accomplish the goal in this paper. Next, it is explained in detail how a Spanish translated large-scale dataset was obtained from the adapted SQuAD dataset for QG task.

First, the original sentences and paragraphs in adapted SQuAD dataset for QG are tokenized. So, the reverse process has been executed, in order to get the original paragraphs and sentences, the next steps were done:

- All -lrb- and -rrb- were replaced by '(' and ')' respectively,
- The `` and '' were replaced by double quotes, " and ".
- All '--' were replaced by a single '-'.
- All the first letter of each sentence and paragraph were capitalized
- All commas (,), dots (.), semicolons(;), and colons(:) were put after the word the follow, to eliminate spaces.
- All possessive nouns ('s) where set correctly by eliminating spaces between the word, apostrofe, and letter 's'.

An example of such procedure could be seen in Table II, where the des-tokenized sentence is more human readable and in order to get high accuracy on the translated data by using Google Translation API.

TABLE II
DES-TOKENIZING OF SENTENCES AND PARAGRAPHS

Original Tokenized Sentence	**Example 1** the american football conference -lrb- afc -rrb- champion denver broncos defeated the national football conference -lrb- nfc -rrb- champion carolina panthers 24 -- 10 to earn their third super bowl title . **Example 2** jaime weston , the league 's vice president of brand and creative , explained that a primary reason for the change was the difficulty of designing an aesthetically pleasing logo with the letter `` l '' using the standardized logo template introduced at super bowl xlv .
Des-tokenized Sentence	**Example 1** The american football conference (afc) champion denver broncos defeated the national football conference (nfc) champion carolina panthers 24-10 to earn their third super bowl title. **Example 2** Jaime weston, the league's vice president of brand and creative, explained that a primary reason for the change was the difficulty of designing an aesthetically pleasing logo with the letter (l) using the standardized logo template introduced at super bowl xlv.

Second, Table I describes the amount of sentences and paragraphs contained in each file, to do a work more efficiently, this data has been separated in small files, each small file contained 500 lines of sentences or paragraphs. In this way the responses from Google Translation API were faster, accurate and more reliable. Third, Once all the paragraphs and sentences on each small file were translated, they have been joined all again to get a new Spanish translated dataset for QG from English adapted SQuAD dataset for QG task. The results of the translated des-tokenized data could be seen in Table III.

TABLE III
DES-TOKENIZING OF SENTENCES AND PARAGRAPHS

Des-tokenized Sentence	**Example 1** The american football conference (afc) champion denver broncos defeated the national football conference (nfc) champion carolina panthers 24-10 to earn their third super bowl title. **Example 2** Jaime weston, the league's vice president of brand and creative, explained that a primary reason for the change was the difficulty of designing an aesthetically pleasing logo with the letter (l) using the standardized logo template introduced at super bowl xlv.
Translated Spanish Sentence by Google Translation API	**Example 1** El campeón de la conferencia de fútbol americano (afc) denver broncos derrotó a la campeona de carolina panteras de la conferencia nacional de fútbol (nfc) 24-10 para ganar su tercer título del Super Bowl. **Example 2** Jaime Weston, vicepresidente de marca y creativo de la liga, explicó que la razón principal del cambio fue la dificultad de diseñar un logotipo estéticamente agradable con la letra (l) que utiliza la plantilla de logotipo estandarizada presentada en el Super Bowl xlv.

Fourth, in order to train a new model using this new Spanish translated dataset for QG, a tokenizing step is required. In order to tokenize all the sentences and paragraphs, Stanford Core NLP [41] has been used, then all the paragraphs and sentences were converted to lowercase. The resulting tokenized Spanish Translated data is shown in Table IV.

TABLE IV
TOKENIZED SPANISH TRANSLATED DATA USING STANFORD CORE NLP

Translated Spanish Sentence by Google Translation API	**Example 1** El campeón de la conferencia de fútbol americano (afc) denver broncos derrotó a la campeona de carolina panteras de la conferencia nacional de fútbol (nfc) 24-10 para ganar su tercer título del Super Bowl. **Example 2** Jaime Weston, vicepresidente de marca y creativo de la liga, explicó que la razón principal del cambio fue la dificultad de diseñar un logotipo estéticamente agradable con la letra (l) que utiliza la plantilla de logotipo estandarizada presentada en el Super Bowl xlv.
Tokenized sentence in lowercase	**Example 1** el campeón de la conferencia de fútbol americano -lrb- afc -rrb- denver broncos derrotó a la campeona de carolina panteras de la conferencia nacional de fútbol -lrb- nfc -rrb- 24 -- 10 para ganar su tercer título del super bowl . **Example 2** jaime weston , vicepresidente de marca y creativo de la liga , explicó que la razón principal del cambio fue la dificultad de diseñar un logotipo estéticamente agradable con la letra -lrb- l -rrb- que utiliza la plantilla de logotipo estandarizada presentada en el super bowl xlv .

Until now, the Spanish translated dataset has been set up and it is ready to train a model to generate questions from sentences using neural networks in Spanish language. Note that 'denver broncos' could be less familiar than 'jaime watson' for Spanish language even though both are in English.

B. *OpenNMT Machine Translation*

OpenNMT [20] is a toolkit for Neural Machine Translation (NMT), which is largely used for NLP community. In this paper, it has been used to compare the translations done by OpenNMT and Google Translation API on adapted SQuAD for QG.

This toolkit requires a parallel corpus to train Machine Translation models, which means two pair of sentences from two different languages, it has been used the European Parliament English-Spanish parallel corpus (europarlv7) [33].

The europarlv7 dataset has 1 965 734 English-Spanish pair of sentences. It has been splitted randomly into a training set (80%), a development set (10%), and a test set (10%). The sentences where tokenized for both Spanish and English as shown in Table V.

TABLE V
EUROPARLV7 PARALLEL CORPUS ENGLISH-SPANISH TOKENIZED SENTENCES

Original Sentences in parallel corpus Europarlv7	**English Sentence** In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union. **Spanish Sentence** A la espera de que se produzca, de acuerdo con muchos colegas que me lo han pedido, pido que hagamos un minuto de silencio en memoria de todas las víctimas de las tormentas, en los distintos países de la Unión Europea afectados.
Tokenized Sentences from parallel corpus Europarlv7	**English Sentence** in the meantime , i should like to observe a minute ' s silence , as a number of members have requested , on behalf of all the victims concerned , particularly those of the terrible storms , in the various countries of the european union . **Spanish Sentence** a la espera de que se produzca , de acuerdo con muchos colegas que me lo han pedido , pido que hagamos un minuto de silencio en memoria de todas las víctimas de las tormentas , en los distintos países de la unión europea afectados .

Once the data has been set up to train a Machine Translation model that could translate English sentences into Spanish sentences. The results obtained by OpenNMT on europarlv7 parallel corpus are shown in Table VI. The results seems to be promising since the data in the test set are in the same context as the training and development dataset.

TABLE VI
RESULTS OBTAINED ON TEST SET OF EUROPARLV7

BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
0.31	0.23	0.11	0.09	0.14

This trained model has been used to translate adapted SQuAD for QG dataset to be compared against Google Translation API. The results are shown in Table VII. It is clearly seen that the trained model have not produced accurate readable sentences as it was produced in testing set in Table VI.

TABLE VII
Translation on Adapted SQuAD Dataset for QG by Using OpenNMT Trained Model

Des-tokenized Sentence from Adapted SQuAD Dataset for QG Task	**Example 1** The american football conference (afc) champion denver broncos defeated the national football conference (nfc) champion carolina panthers 24-10 to earn their third super bowl title. **Example 2** Jaime weston, the league's vice president of brand and creative, explained that a primary reason for the change was the difficulty of designing an aesthetically pleasing logo with the letter (l) using the standardized logo template introduced at super bowl xlv.
Translated Spanish Sentence by Trained Model OpenNMT on europarlv7 parallel corpus	**Example 1** Los caballos salvajes de Denver del campeón de la conferencia de fútbol americano (afc) derrotaron las panteras de Carolina del campeón de la conferencia de fútbol nacional (nfc) 24-10 para ganar su tercer título del cuenco estupendo **Example 2** El weston de Jaime, el vicepresidente el de la liga de la marca y creativo, explicados que una razón primaria del cambio era la dificultad de diseñar un logotipo estético satisfecho con la letra (l) usando la plantilla estandardizada del logotipo introdujo en el xlv del cuenco estupendo.

V. Evaluation of the New Translated Spanish Data from Adapted SQuAD Dataset for QG

Since the results got by OpenNMT trained model using European Parliament dataset (europarlv7) [33] are less coherent compared to Google Translation API results, as shown in Table VII. This section only presents the evaluation of the new Spanish translated dataset from adapted SQuAD dataset for QG by using Google Translation API.

A. Metrics for Evaluation of Translated Data

Metrics used for evaluation of results in NLP tasks are mainly of two kind. On one hand, there are automatic evaluation metrics such as BLEU [42], METEOR [43], and ROUGE[44]. On the other hand, Human Evaluation is used largely because it is necessary, provides a better understanding and measures fluency and adequacy of the sentences as suggested by [8], [21] [30] even though it is time consuming to evaluate thousands of sentences.

Bilingual Evaluation Understudy commonly known as BLEU [42]. In this metric the precision and recall are approximated by modified n-gram precision and best match length, respectively.

Metric for Evaluation for Translation with Explicit Ordering commonly known as METEOR [43]. This metric claims to be better correlation with human judgements.

Recall Oriented Understudy for Gisting Evaluation commonly known as ROUGE [44]. This metric is mostly used for summary evaluation and it is entirely based on the Longest Common Subsequence (LCS).

As suggested by [30], human evaluation was done according to fluency and adequacy scales prepared by [30] since it provides a better understanding and they are easy to apply, such scales are shown in Table VIII.

TABLE VIII
Fluency and Adequacy Scales (Reyes Ayala, 2018)

Scale	Fluency	Adequacy
5	**Flawless**: Translated text fully conforms to rules of the language and is consistent with evaluator’s use of native language.	**All**: Completely match the meaning of at least one of the reference translations. All parts are correctly translated
4	**Good**: Translated text conforms to rules of language to some extent and is partly consistent with the evaluators use of native language	**Most**: Most parts are correctly translated
3	**Non-native**: Translated text is understandable but not consistent with the evaluators use of native language	**Much**: Half or more is correctly translated, but fewer than Most
2	**Disfluent**: Translated text is barely understandable	**Little**: Less than half are correctly translated, some important concepts are not correctly translated
1	**Incomprehensible**: Translated text is totally beyond understanding	**None**: Totally different in meaning from the references

B. Evaluation of New Spanish Translated Dataset from Adapted SQuAD for QG

The evaluation of the new Spanish translated dataset from adapted SQuAD dataset for QG task has been done using the metrics explained in section V, such as BLEU, METEOR, ROUGE and Human Evaluation for fluency and adequacy.

From the adapted SQuAD dataset for QG that contains 92 931 sentences, it has been taken randomly 979 sentences, that represent the sample size with a margin error of 3.12 and confidence level of 95%.

These 979 sentences were translated for two human Spanish native speaker experts that know and understand English language. The result of translations between Google Translation API and human experts are shown in Table IX.

TABLE IX
979 Translated Sentences by Google Translation API and Human Experts

Original Sentence: The annual bookstore basketball tournament is the largest outdoor five-on-five tournament in the world with over 700 teams participating each year, while the notre dame men's boxing club hosts the annual bengal bouts tournament that raises money for the holy cross missions in bangladesh. **Google Translation API**: El torneo anual de baloncesto de la librería es el mayor torneo al aire libre de cinco a cinco del mundo, con más de 700 equipos participando cada año, mientras que el club de boxeo de notre dame para hombres organiza el torneo anual de combates de bengala que recauda dinero para las misiones de la Santa Cruz en Bangladesh. **Human Expert**:
Original Sentence:

El torneo anual de baloncesto de bibliotecas es el torneo al aire libre de cinco a cinco más grande del mundo con más de 700 equipos que participan cada año, mientras que el club de boxeo para hombres de notre dame organiza el torneo anual de combates de bengala que recauda dinero para las misiones de la Santa Cruz en Bangladesh.
The old college building has become one of two seminaries on campus run by the congregation of holy cross. **Google Translation API**: El antiguo edificio de la universidad se ha convertido en uno de los dos seminarios en el campus dirigido por la congregación de la Santa Cruz. **Human Expert**: El antiguo edificio de la universidad se ha convertido en uno de los dos seminarios en el campus dirigido por la congregación de la Santa Cruz.

The results of the evaluation of the 979 sentences using the established metrics, including the human evaluation are shown in Table X.

TABLE X
Results using Evaluation Metrics on 979 Random Sentences from Spanish Translated Data from Adapted SQuAd for QG.

Metrics	Average
BLEU-1	0.39
ROUGE	0.35
METEOR	0.17
FLUENCY	Non-native
ADEQUACY	Much

Table X shows that results for fluency and adequacy are Non-native and Much respectively, according to the scale. Which means the Spanish translated data from the adapted SQuAD for QG task is understandable and readable.

VI. State-of-the-art Neural Network Model for Question Generation (QG)

Now, the new Spanish translated dataset has been evaluated and its acceptance has been validated by human experts, this translated dataset could be used to train a model for automatic question generation for Spanish language. Next , this section describes the neural network model used to train the new Spanish translated dataset, which is the same used in [4] and also describes some settings done for word embeddings, which was an important part to train the question generation model.

A. Recurrent Neural Network

The model used in [4] has been adapted from OpenNMT [20] and it is inspired by the way in which a human would ask a question, paying attention to certain parts of the sentence and also associating context information. The model in [4] uses the most famous neural network in NLP, Recurrent Neural Network (RNN) where the architecture allow to exhibit temporal dynamic behaviour for a time sequence, which means it can use their internal memory to process sequences of inputs sentences [46], [47].

B. Settings for Training the Question Generation Model

In order to use the sequence-to-sequence model presented in [4]. Besides the dataset, it is required the vector of words in the language that it is being trained. (Du et. al) [4] uses Glove word embedding of 300 dimensions, which has been pre-trained by [23], but that is the set for English language.

So, for Spanish language it is needed a similar approach, it has been used several pre-trained embedding words, Vect2Word [22], Glove [23], and FastText[45]. these word embeddings were trained using SBWCE [35] corpus. the dimensions of such models are of 300, the same as in [23]. Table XI shows the different pre-trained models for word embedding used in this paper.

TABLE XI
Word Embedding vectors used for Spanish Language

N°	Embedding	Corpus	Algorithm

1	dimension = 300
	vectors = 855380
		SBWCE	FastText

2	dimension = 300
	vectors = 855380
		SBWCE	Glove

3	dimension = 300
	vectors = 985667
		Wikipedia Spanish Dump	FastText

4	dimension = 300
	vectors = 1000653
		SBWCE	Vect2Word

VII. Evaluation of Question Generation (QG) Model Trained with Spanish Translated Data

The evaluation of the trained Question Generation Model for Spanish language has been carried out by using the same metrics described in this paper, BLEU, ROUGE, METEOR and the human evaluation for the generated questions from the sentences in the test set (10%) established before.

Table XII shows the result of questions generated by the trained model. as it is shown the expected questions are very similar to the generated question by the model.

TABLE XII
Generation of Questions by the Trained Model on Spanish Language

Original Sentence: Notable athletes include swimmer sharron davies, diver tom daley , dancer wayne sleep, and footballer trevor francis. **Translated Sentence** Entre los atletas notables se encuentran el nadador Sharron Davies, el buceador Tom Daley, el bailarín Wayne Sleep y el futbolista trevor francis. **Expected Question to be Generated:** ¿Cuál es la ocupación de trevor francis? **Question Generated by the Trained Model:** ¿cuál es el nombre de el nadador sleep?
Original Sentence: Father joseph carrier, c.s.c. was director of the science museum and the library and professor of chemistry and physics until 1874. **Translated Sentence** Padre José, portador, c.s.c. Fue director del museo de ciencias y de la biblioteca y profesor de química y física hasta 1874. **Expected Question to be Generated:** ¿Qué persona fue el director del museo de ciencias en Notre Dame a fines del siglo XIX? **Question Generated by the Trained Model:** ¿Quién es el director de el museo de ciencias y física hasta 1874?
Original Sentence: Canadian politician and legal scholar chris axworthy hails from plymouth. **Translated Sentence** El político canadiense y erudito legal Chris Axworthy es oriundo de Plymouth.

Expected Question to be Generated:
¿Cuál es la nacionalidad actual del ex residente de plymouth chris axworthy?
Question Generated by the Trained Model:
¿Cuál es el nombre de el gobierno canadiense?

The results shown in Table XII were generated using the word embedding number 4 from Table XI since this pre-trained word embedding got high scores among the other word embedding models.

The results of evaluation using automatic metrics BLEU, ROUGE, METEOR are shown in Table XIII. Unlike automatic evaluation metrics, Adequacy and fluency where tested on 869 interrogative sentences from the test set. It also shows the increment on each metric compared to original English work [4].

TABLE XIII
Results using Evaluation Metrics on Spanish Translated data

Metrics	English Adapted SQuAD for QA	Spanish Translated Data	Increment
BLEU-1	0.39	0.45	15 %
BLEU-2	0.21	0.27	29 %
BLEU-3	0.13	0.19	46 %
BLEU-4	0.08	0.14	75 %
ROUGE	0.35	0.42	20 %
METEOR	0.13	0.18	38 %
FLUENCY	-	Non-native	-
ADEQUACY	-	Much	-

VIII. Conclusions and Future Work

The approach used in this paper, generating automatic questions from paragraphs and sentences using translated large-scale dataset to train an automatic question generation model for Spanish language has shown that the quality of the data used is very important as remarked in [12]. Since the quality of the translated data obtained using OpenNMT is far away from the ratio obtained by Google Translation API, it has been discarded from evaluations. It has been clearly shown that a huge parallel corpus is needed to translate all the sentences in adapted SQuAD for QG task dataset.

Google Translation API has shown to be the best machine translation using adapted SQuAD dataset for QG task since it has translated the sentences and paragraphs with a moderate high acceptance ratio by human evaluation, but the adapted SQuAD dataset for QG task has many repeated sentences, and the questions that are related to sentences and paragraphs could be find along the sentence and paragraphs where the question was defined, which means, it does not required reasoning in case the question needs to be answered.

The results have shown that the translated data outperformed the previous work done in the English language by 37% in average on all metrics. This result shows how repeated are the sentences in adapted SQuAD dataset and the importance of the word embedding since it has been the only factor that made the results changed on the training process.

It could also be seen that neural network models trained by specific task using RNN is independent of the language used. In this paper, Spanish translated data has been used and the results are still considered as good as on the original language.

In the near future, it is expected to build a large-scale dataset for Spanish Question Generation language for educational purposes. In this escenario, the answers to the questions are required to be difficult to be found along the sentence, which means, reasoning is required to be able to answer the questions. In this way the intention of the authors of this paper are to establish a new baseline for Question Generation (QG) task for Spanish language using neural networks with a real Spanish high quality dataset.

Acknowledgment

This research is part of the project "AQNLP: Software de Apoyo a la Comprensión Lectora de Estudiantes en Etapa Escolar, Mediante la Generación de Preguntas Automáticas desde Libros de Texto Utilizando Tecnología NLP", under a grant contract N° TP-053-2018-UNSA, which is supported by Universidad Nacional de San Agustín de Arequipa. It is also a great opportunity to thank CiTeSoft-UNSA for the great support and the great environment provided during this research.

References

[1] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science, vol. 349, no. 6245, pp. 261–266, 2015.

[2] R. A. Calvo, D. N. Milne, M. S. Hussain, and H. Christensen, "Natural language processing in mental health applications using non-clinical texts," Natural Language Engineering, vol. 23, no. 5, pp. 649–685, 2017.

[3] R. E. Jack, C. Crivelli, and T. Wheatley, "Data-driven methods to diversify knowledge of human psychology," Trends in cognitive sciences, vol. 22, no. 1, pp. 1–5, 2018.

[4] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), (Vancouver, Canada), pp. 1342–1352, Association for Computational Linguistics, July 2017.

[5] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, (Copenhagen, Denmark), pp. 866–874, Association for Computational Linguistics, September 2017.

[6] X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Zhang, S. Subramanian, and A. Trischler, "Machine comprehension by text-to-text neural question generation," in Proceedings of the 2nd Workshop on Representation Learning for NLP, (Vancouver, Canada), pp. 15–25, Association for Computational Linguistics, August 2017.

[7] V. Kumar, K. Boorla, Y. Meena, G. Ramakrishnan, and Y.-F. Li, "Automating reading comprehension by generating question and answer pairs," arXiv preprint arXiv:1803.03664, 2018.

[8] M. Heilman and N. A. Smith, "Good question! statistical ranking for question generation," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (Los Angeles, California), pp. 609–617, Association for Computational Linguistics, June 2010.

[9] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[10] M. Malinowski, M. Rohrbach, and M. Fritz, “Ask your neurons: A neural based approach to answering questions about images,” in Proceedings of the IEEE international conference on computer vision, pp. 1–9, 2015

[11] L. Ma, Z. Lu, and H. Li, “Learning to answer questions from image using convolutional neural network.,” in AAAI, vol. 3, p. 16, 2016.

[12] C. H. Guanliang Chen, Jie Yang and G.-J. Houben, “Learningq: A large-scale dataset for educational question generation,” 2018.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, (Austin, Texas), pp. 2383–2392, Association for Computational Linguistics, November 2016.

[14] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, “Ms marco: A human generated machine reading comprehension dataset,” arXiv preprint arXiv:1611.09268, 2016.

[15] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, and T. Mikolov, “Towards ai-complete question answering: A set of prerequisite toy tasks,” arXiv preprint arXiv:1502.05698, 2015.

[16] E. Choi, H. He, M. Iyyer, M. Yatskar, W.-t. Yih, Y. Choi, P. Liang, and L. Zettlemoyer, “Quac: Question answering in context,” arXiv preprint arXiv:1808.07036, 2018.

[17] S. Reddy, D. Chen, and C. D. Manning, “Coqa: A conversational question answering challenge,” arXiv preprint arXiv:1808.07042, 2018.

[18] Y. Yang, W.-t. Yih, and C. Meek, “Wikiqa: A challenge dataset for open domain question answering,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, (Lisbon, Portugal), pp. 2013– 2018, Association for Computational Linguistics, September 2015.

[19] A. Paitamala and D. Augusto, “Generación de preguntas sobre un texto,” 2016.

[20] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, “Opennmt: Open source toolkit for neural machine translation,” in Proc. ACL, 2017.

[21] V. Kumar, K. Boorla, Y. Meena, G. Ramakrishnan, and Y.-F. Li, “Automating reading comprehension by generating question and answer pairs,” arXiv preprint arXiv:1803.03664, 2018.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[23] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, October 2014.

[24] T. Baghaee, Automatic Neural Question Generation using Community-based Question Answering Systems. PhD thesis, Lethbridge, Alta.: University of Lethbridge, Dept. of Mathematics and Computer Sciences, 2017.

[25] P. Schäuble and P. Sheridan, “Cross-language information retrieval (clir) track overview,” in In Proceedings of the Sixth Text Retrieval Conference (TREC-6, Citeseer, 1997.

[26] M. Braschler, J. Krause, C. Peters, and P. Schäuble, “Cross-language information retrieval (clir) track overview,” in TREC, 1999

[27] C. Peters, Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000, Revised Papers, vol. 2069. Springer, 2003.

[28] Cámara, E. (2019). *TASS 2017 @SEPLN*. [online] Sepln.org. Available at: http://www.sepln.org/workshops/tass/2017/ [Accessed 1 Feb. 2019].

[29] Sepln.org. (2019). *sepln*. [online] Available at: http://www.sepln.org/ [Accessed 1 Feb. 2019].

[30] B. Reyes Ayala, R. Knudson, J. Chen, G. Cao, and X. Wang, “Metadata records machine translation combining multi-engine outputs with limited parallel data,” Journal of the Association for Information Science and Technology, vol. 69, no. 1, pp. 47–59, 2018.

[31] J. Camacho-Collados, M. T. Pilehvar, N. Collier, and R. Navigli, “Semeval 2017 task 2: Multilingual and cross-lingual semantic word similarity,” in Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), (Vancouver, Canada), pp. 15–26, Association for Computational Linguistics, August 2017.

[32] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, “Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation,” arXiv preprint arXiv:1708.00055, 2017.

[33] Statmt.org. (2019). *Europarl Parallel Corpus*. [online] Available at: http://www.statmt.org/europarl/ [Accessed 1 Feb. 2019].

[34] F. Enriquez, F. Cruz, F. Javier Ortega, and J. A. Troyano, “Spanish-english similarity through word embeddings,” PROCESAMIENTO DEL LENGUAJE NATURAL, no. 59, pp. 31–38, 2017.

[35] C. Cardellino, “Spanish Billion Words Corpus and Embeddings,” March 2016.

[36] M. Nadhianti, “An analysis of accuracy level of google translate in english bahasa indonesia and bahasa indonesia-english translations,” Sastra Inggris Quill, vol. 5, no. 4, pp. 296–303, 2016.

[37] H. Ghasemi and M. Hashemian, “A comparative study of"google translate"translations: An error analysis of english-to-persian and persian-to-english translations.,” English Language Teaching, vol. 9, no. 3, pp. 13–17, 2016.

[38] L.-R. Precup-Stiegelbauer, “Automatic translations versus human translations in nowadays world,” Procedia-Social and Behavioral Sciences, vol. 70, pp. 1768–1777, 2013.

[39] M. Saffari, S. Sajjadi, and M. Mohammadi, “Evaluation of machine translation (google translate vs. bing translator) from english into persian across academic fields,” MODERN JOURNAL OF LANGUAGE TEACHING METHODS, vol. 7, no. 8, pp. 429–442, 2017.

[40] M. Groves and K. Mundt, “Friend or foe? google translate in language for academic purposes,” English for Specific Purposes, vol. 37, pp. 112–121, 2015.

[41] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit,” in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55–60, 2014.

[42] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311–318, Association for Computational Linguistics, 2002.

[43] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72, 2005.

[44] E. Reiter and A. Belz, “An investigation into the validity of some metrics for automatically evaluating natural language generation systems,” Computational Linguistics, vol. 35, no. 4, pp. 529–558, 2009.

[45] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” Transactions of the Association for Computational Linguistics, vol. 5, pp. 135–146, 2017.

[46] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in Advances in neural information processing systems, pp. 577–585, 2015.

[47] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.