Neural Machine Translation tool from Spanish to English in the medical domain.

Gordillo Lucas, Ariel Santiago

Departamento de Ciencias de la Computación

Carrera de Software

Artículo académico, previo a la obtención del título de Ingeniero en Software

Msc. Uyaguari Uyaguari, Alvaro Danilo

Latacunga

18 de agosto de 2022

Neural Machine Translation tool from Spanish to English in the medical domain

Ariel Gordillo
Departamento de Ciencias de la Computación
Universidad de las Fuerzas Armadas ESPE
Latacunga
Cotopaxi, Ecuador
asgordillo1@espe.edu.ec

Alvaro Uyaguari
Departamento de Ciencias de la Computación
Universidad de las Fuerzas Armadas ESPE
Latacunga
Cotopaxi, Ecuador
aduyaguari@espe.edu.ec

Lucas Garces
Departamento de Ciencias de la Computación
Universidad de las Fuerzas Armadas ESPE
Latacunga
Cotopaxi, Ecuador
lrgarces@espe.edu.ec

***Abstract* - In Natural Language Processing (NLP), the scarcity of linguistic resources (labeled corpus, parallel corpus, pre-trained models, etc.) can lead to poor performance when applying machine learning models, however, this can be solved by applying cross-lingual approaches (machine translation, word alignment, multilingual embedding, multilingual embedding, etc.), which is a paradigm for transferring knowledge from one language with resources to another language with fewer resources. In the medical domain, there are also few resources in Spanish compared to English, due to economic, legal, and ethical issues. In this regard, there is little evidence of evaluation and optimization of machine translations from Spanish to English in the medical domain. For this purpose, a neural machine translation tool with an induced word alignment is generated in this research, on which different optimization parameters have been experimented with and applying various parallel corpora within the medical domain, as reference results with the corpora EMA with 15 epochs, a BLUE of 88.55 in English-Spanish and Scielo Spanish - English with 25 epochs, a BLEU of 53.74 , being a differential in evaluation results to convolutional translators and even greatly outperforming the pre-trained Fairseq results.**

Keywords - machine translation; Cross-Lingual; word alignment; medical domain; natural language processing; NLP.

I. INTRODUCTION

Currently, the scarcity of linguistic resources in Spanish (labeled corpora, parallel corpora, pretrained models, etc.) in the medical domain is due to ethical, economic and legal reasons [1], thus limiting the creation of new machine learning models in the medical context. For this reason, we propose to increase biomedical resources by means of a Neural Machine Translation (NMT) tool with an induced word alignment for knowledge extraction from the English language, where linguistic resources are greater, to the Spanish language. This, in order to later implement an identifier of medical entities, that through the NMT where knowledge is extracted with systems such as Unified Medical Language System (UMLS) or database in English that are within the medical domain and with the help of word alignment, to bring these labels back to the language of origin, Spanish. Consequently, once these resources are generated, use and generate a system of prediction of diseases, diagnoses and medical treatments, this process is described in the "Fig 1".

For this purpose our research hypothesis therefore is that if we create a Spanish to English translator in the medical domain then we increase linguistic resources in this language. To validate this hypothesis, we compare the effectiveness of recognition of medical entities from the translations generated by the commercial translator with our translations focused in the medical domain. In the end, it is evident that the best model obtains a BLEU of 88.55, making the translations more accurate and allows, also, through a comparison with a commercial translator, a greater number of biomedical entities were recognized.

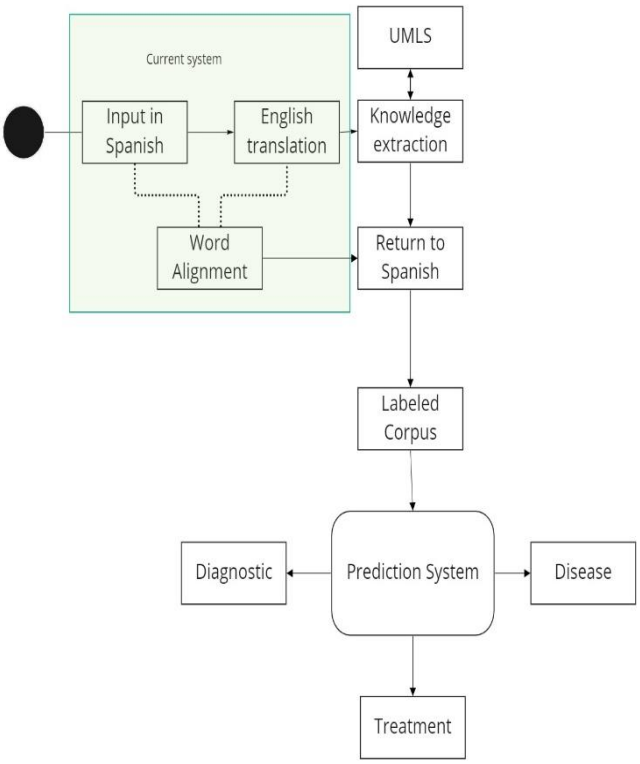

Fig 1. System projection

To carry out the creation of our translator, a set of tools was identified, of which the Fairseq toolkit was used for data preprocessing and training. Also, parallel corpora from Spanish to English oriented in the medical domain were found. Similarly, the Transformer model was implemented, which according to [2] works well for language pairs, allowing according to [3] huge performance improvements in multiple natural language processing tasks.

Next, we present the sections that make up the article, where we describe the process we followed to carry out our research, which are: i) Related works: here we describe the

current state of the art of the main tools and models that support the NMT. ii) Implementation: here we present the phases that we carried out for the creation of the translation model such as tool selection, corpus selection, pre-screening, inference and the translator's validation process. iii) Results: we analyze the validation and the score produced by the translator, as well as the analysis of a comparison of our translator with a commercial translator. iv) Conclusions and Future work: Finally we make conclusions and describe the future work that can be carried out from the use of our automatic translation tool.

II. Related work

A. Neural machine translation

In the field of machine translation of written text from one natural language to another, has undergone a major paradigm shift in recent years. In 2010 [4] presented a translation system where connectionist n-grams are integrated in the decoding stage, because this model gave the best results up to that date.

In 2014 [5] propose a new architecture with a recurrent encoder-decoder neural network that is able to learn the mapping from one sequence to another sequence.

In 2016 [6] presented as a goal to build a single neural network that can be jointly tuned to maximize translation performance. In 2017 [7] propose a simple solution to use a single multilingual NMT model, their approach allows systems to use a single model allowing benchmarks such as WMT'14, achieve comparable performance for English - French and outperforms more advanced results for English – German. This same year [8] instead of using convolutional networks, implements stacked self-attention layers, significantly improving the state of the art of machine translation and language modeling, improving also the training speed.

However, in 2019 [9] propose a new seq2seq translation architecture to highlight the importance of sequential dependencies in contexts for sequential recommendation. Being this model some of the most popular and with better results in translations, due to them in Table 1. are shown the tools that currently provide seq2seq translations and their impact in recent years.

In addition, it can be evidenced that the scores of Spanish to English translators are lower than the ones we present in this research, as for example: in 2017 a multilanguage translator from Spanish to English is implemented, generating a BLEU of 29.7, implementing machine translation technologies [10], however, in 2020 biomedical translation tasks are implemented with recurrent neural networks using the EMA corpus and reaching a BLEU of 0.541 points [11].

In 2019 [12] proposes a model evaluation for translation where it generates a BLEU of 56.47, meanwhile, translators oriented to the medical domain, we can find a biomedical translator with word alignment of WMT 2021 from Spanish to English [13], where its highest Bleu is 0.5382 being the highest among a set of test parameters..

B. Cross-Lingual

Cross-Lingual in Natural Language Processing (NLP) is an important alternative in the development of a system based on one or more languages for which few resources are available according to [14]. Cross-lingual is the solution to solve this lack of data in resource-poor languages. It mainly involves using annotated data from other languages to build and define new NLP models as mentioned by [15].

In this way, cross-lingual can help to create intelligent systems in languages where it was not possible before and improve their performance. Usually, some kind of multilingual resources or technology (parallel corpora, multilingual distributional representations, word alignment, etc.) is used to solve the difference between languages, without these resources the gap between language pairs may be too large for machine learning methods [15].

C. Word Alignment

According to [16] word alignment is defined as the detection of the corresponding alignment between words in parallel sentences translated from one another.

Nowadays there are several models and tools that implement word alignment, such as: i) GIZZA++ [17] a word alignment tool that facilitates the development of machine translation systems. ii) Berkley Aligner [18] which allows a supervised and unsupervised approach to align words in parallel corpora. In our research, word alignment is aimed at the future projection of medical entities from English to Spanish.

III. Implementation

For the implementation of our tool we first conducted a review of the state of the art for the choice of the tool that will allow us to preprocess and train the data, then we conducted a search to determine the parallel corpora that are within the medical domain, once selected the environment and the corpora, we proceed to preprocess and train, at this point we induce the alignment of words and determine the inference of the translations and finally we perform a validation process of the generated translation, through a comparison between our tool and a commercial translator. As shown in “Fig 2.” the phases of the implementation of our tool.

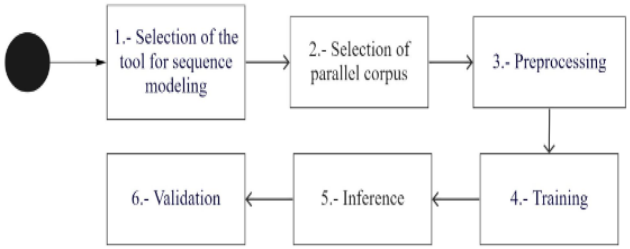

Fig 2. Implementation Flow

A. Phase 1: Selection of the tool for sequence modeling

In this phase, as a first step, a review of the state of the art was carried out, where the most widely used projects for research in sequential models to date were identified, which are shown in Table I.

Based on the review, it was concluded that Fairseq, being the project with the highest number of bibliographic citations, was chosen as the basis for the preprocessing and training of the model for the NMT.

Fairseq is an open source sequence modeling toolkit which allows researchers and developers to train custom models for translation, summarization, language modeling and other text generation tasks [19]. This toolkit is based on PyTorch and supports distributed training across multiple GPUs. Also, the use of the Transformer architecture was detected in the kit

chosen for NMT. This, according to [14] is based on an encoder stack and a decoder stack, where the encoder maps an input sequence of tokens to a sequence of continuous vector representations, where the decoder generates an output sequence of symbols, one element at a time. Making both models and tasks compatible and running with the transformer network.

TABLE I. SEQUENCE MODELING TOOLS.

Toolkit	Sponsor	Google Scholar citations	Last update	Language	Implement Word Alignment
OpenSeq2Seq	NVIDIA	31	2 years ago	Python	No
OpenNMT	Harvard NLP and SYSTRAN	1645 Version 2017 20 Version 2020	3 months ago	PyTorch inits latest version	Yes
Seq2SeqPy	IDEX Universite´ Grenoble Alpes	3	1 year ago	PyTorch	No
Sockeye	Amazon	201 Version 1 7 Version 2	2 months ago	PyTorch	Yes
Fairseq	Facebook	1439	A few days ago	PyTorch	Yes

B. Phase 2: Selection of parallel corpus

The parallel corpora used in this research are as follows:

- SciELO (Scientific Electronic Library Online): According to [20] this corpus gathers electronic publications of articles and full texts of scientific journals from Latin America, South Africa and Spain. It is currently present in 15 countries and is supported by the São Paulo Research Foundation (FAPESP) and the Brazilian National Council for Scientific and Technological Development (BIREME).
- EMEA: EMEA is a corpus containing biomedical documents belonging to the European Medicines Agency (EMEA). The corpus [21] includes documents related to medicinal products and their translations into 22 official languages of the European Union.

C. Phase 3: Preprocessing

Fairseq handles files in the ssh language, which allows the development of scripts for data processing. By means of .sh files a division of the corpus is made, thus generating a set of parallel files: train, test and valid, each one in Spanish and English respectively.

The Moses library is used as a tokenizer for each of the files, as well as the subword NMT where the subwords are identified. Once the preliminary data is obtained, it is preprocessed using fairseq-preprocess, where the en-es data set is binarized.

D. Phase 4: Training

For this phase, fairseq-train is implemented for model training, using TranslationTask. The Transformer architecture "Fig 3." was also used for the NMT training focused on the fairseq implementation. For which the transformer network is based on an encoder stack and a decoder stack, where the encoder maps an input sequence of tokens to a sequence of continuous vector representations, the decoder generates an output sequence of symbols, one element at a time.

At each step, the model is autoregressive and consumes the previously generated tokens as additional input when generating the next token. Then, a guide model is specified, in our case it is initialized with a word alignment model TransformerAlignModel, and transformer_wmt_en_de_big align is taken as reference and consequently load-alignments are loaded to enhance the alignment. Subsequently, each epoch generated is stored by default in a file called checkpoints.best, in which the epoch with the best performance is stored, being this the most important file, since it is the model used in the generation of plain text translations.

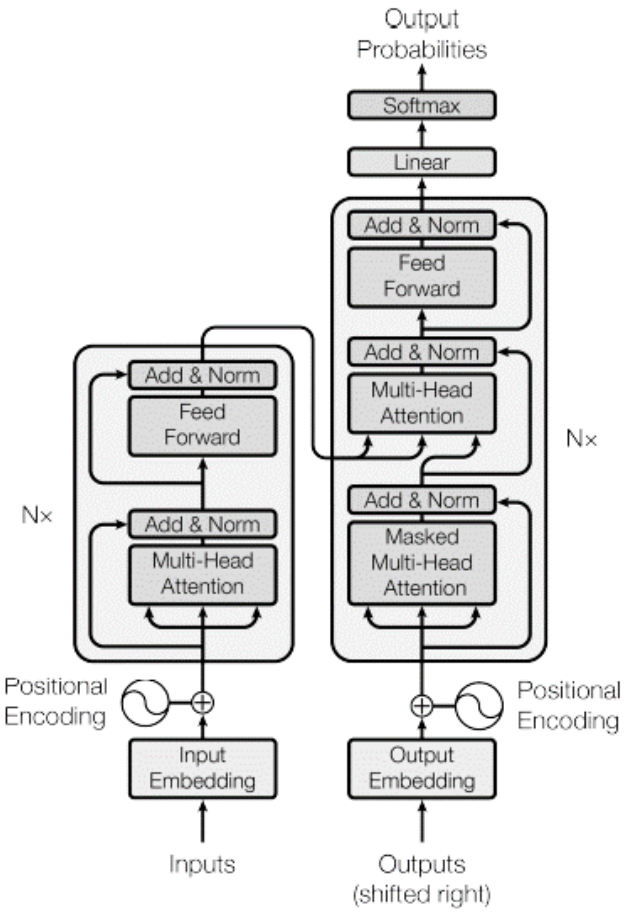

Fig 3. Transformer architectural model

E. Phase 5: Inference

For this phase the parallel files called test are used, since by means of fairseq-generate, supported with the binarized

files, the input language src=en and the output language trg=en are specified, the alignment output is added by print-alignment, and the model, checkpoint.best, is sent with a batch-size 128 and beam 5, as the most important in this step.

The following characteristics are obtained as output: i) S: the sentence in the source language, ii) T: the sentence in the output language, iii) H: the hypothesis of the tokenized translation, iv) D: the hypothesis of the untokenized translation, v) A: the position of each word with respect to its translation. Finally in this process the BLEU value is obtained, this process has been carried out in each of the corpora.

F. Phase 6: Validation

In this phase the BLEU evaluation metric was used, the evaluation results are detailed in Table II. The table contains the BLEU scores of two models generated from the selected parallel corpora, the scores are calculated from the different generating epochs, taking the increment in each of them.

To validate the increase in linguistic resources, a comparison was made with the Microsoft translator, based on the translations generated and thus produce annotations for the recognition of biomedical entities, taking a sample of ten sentences, the results of this comparison are detailed in Table III. This table shows the number of entities recognized by each translator.

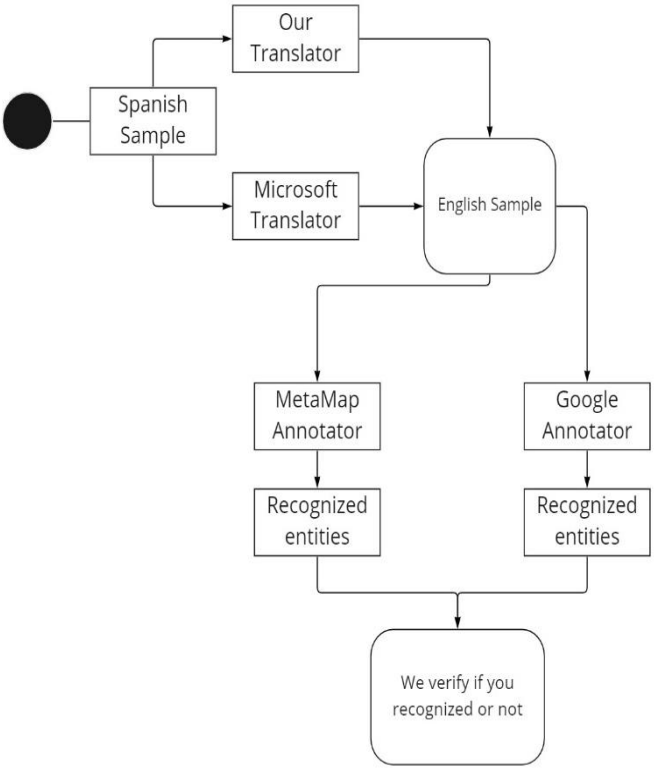

Fig 4. Validation Flow

In "Fig 4." the validation flow of our translator is detailed, which consists of entering the sample sentences and passing them through each translator, the Microsoft translator and our translation tool, obtaining a sample in the English language for each translator, then each of the sentences are sent to the Google annotator and the MetaMap annotator to detect semantic entities. Once semantic entities were detected in the sentences, they were validated and compared with each other, producing that in our translations more entities were detected in the google annotator than the translations generated by the Microsoft translator, while with the MetaMap annotator, the recognized entities were the same for both translators.

IV. RESULTS

As a result of this research and after having carried out the implementation with two parallel corpora and the application of different optimization parameters, the results described in Table II were obtained. As for the BLUE scores of each generated model are presented, being the most outstanding, the model generated by the EMEA corpus, achieving one of the highest scores that have been obtained to date with this toolkit.

TABLE II. BLUE score of our Neural Machine Translator

	EMEA			**SCIELO**		
Epoch	5	8	15	9	15	25
BLEU	78.21	83.61	**88.55**	33.87	45.85	**54.73**
batch-size	128	128	128	128	128	128
beam	5	5	5	5	5	5
train	929044	92904	929044	190351	190351	190351
test	40469	40439	40439	6984	6984	6984
valid	42229	42229	42229	7288	7288	7288

On the other hand, the BLEU of the Scielo corpus outperformed the pre-trained models recommended by Fairseq by several points. Once we knew that the performance of the NMP met our expectations, we verified that unlike other similar translators presented in section II, our translator outperformed with creses with a value of 88.55, generating the best BLEU to date, with only 15 training epochs. Consequently, the efficiency of our translator was tested against a commercial translator such as Microsoft Translator.

For the comparison we used Google's Healthcare Natural Language API [22] annotators and the MetaMap [23] annotator for the recognition of biomedical entities, this from a sample of ten sentences taken from Corpus Gold of the Clef [24].

In the Table III and Table VI show the results of the comparison between the translators, of which a total of 83 medical entities were detected, of which for the Google annotator and with the translations generated by Microsoft only 27 entities were detected, and with our translations we were able to reach 30 entities detected.

Meanwhile with the MetaMap annotator and with the translations generated by Microsoft we were able to detect 52 entities, the same amount that we detected with our translator, given this background we can mention that our translator can recognize up to three more entities than the translations generated by the Microsoft translator.

TABLE III. RESULTS OF THE COMPARISON BETWEEN TRANSLATORS.

Entities recognized by Google Annotator		
Total entities	Microsoft Translator	Our Translator
83	27	**30**

TABLE IV. RESULTS OF THE COMPARISON BETWEEN TRANSLATORS.

Entities recognized by the MetaMap Annotator		
Total entities	Microsoft Translator	Our Translator
83	52	**52**

V. CONCLUSIONS AND FUTURE WORK

Neural machine translation together with the application of Cross-Lingual techniques (machine translation, parallel corpora, word alignment, etc) is a good alternative to solve the shortage of lingustic resources, since knowledge can be extracted from a language with more resources and used in a language with fewer resources, likewise, the selection of the Fairseq tool for model pre-prosessing and training was a great choice since it implements one of the most sophisticated architectures currently available, such as the Tranformer model.

Also our translation model achieved a BLEU of 88.55 showing a high syntactic and contextual accuracy in each of the generated translations, given this we were able to obtain a good Spanish-English translation score within the medical domain.

In addition, through the comparison between translators we were able to detect a greater number of medical entities with respect to the translations generated by the Microsoft translator, thus concluding that our tool equals and surpasses a commercial translator, confirming that we can achieve an increase of biomedical objects, this through neural machine translation and word alignment from Spanish to English in the medical domain.

Future work includes the improvement of the optimization parameters and the inclusion of the recognition of biomedical entities through word alignment, so that the terms labeled in the English language can be projected to the Spanish language, and thus implement them for the training of new models without the need to resort to labeled corpora, because we generate them. Subsequently introduce an automatic prediction system that through training with medical corpuses achieve predictions of diagnoses, treatments and diseases, achieving a robust and self-sufficient system in the medical domain.

REFERENCES

[1] M. Oronoz, K. Gojenola, A. Pérez, A. D. de Ilarraza, y A. Casillas, On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions, J. Biomed. Inform., vol. 56, pp. 318-332, ago. 2015, doi: 10.1016/j.jbi.2015.06.016.

[2] A. C. Dhar, A. Roy, Md. A. Habib, M. A. H. Akhand, y N. Siddique, Transformer Deep Learning Model for Bangla–English Machine Translation, en Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications, Singapore, 2022, pp. 255-265. doi: 10.1007/978-981-16-6332-1_24.

[3] J. Ferrando y M. R. Costa-jussà, Attention Weights in Transformer NMT Fail Aligning Words Between Sequences but Largely Explain Model Predictions. arXiv, 13 de septiembre de 2021. Accedido: 29 de junio de 2022. Disponible en: http://arxiv.org/abs/2109.05853

[4] F. Z. Martínez y M. J. C. Bleda, Traducción automática basada en n-gramas conexionistas, Proces. Leng. Nat., n.o 45, pp. 221-228, 2010.

[5] K. Cho et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, jun. 2014, doi: 10.48550/arXiv.1406.1078.

[6] D. Bahdanau, K. Cho, y Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate. arXiv, 19 de mayo de 2016. Accedido: 29 de junio de 2022. Disponible en: http://arxiv.org/abs/1409.0473

[7] M. Johnson et al., Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, Trans. Assoc. Comput. Linguist., vol. 5, pp. 339-351, oct. 2017, doi: 10.1162/tacl_a_00065.

[8] A. Vaswani et al., Attention is All you Need, en Advances in Neural Information Processing Systems, 2017, vol. 30. Accedido: 27 de junio de 2022. Disponible en: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[9] K. Sun y T. Qian, Seq2seq Translation Model for Sequential Recommendation. arXiv, 14 de enero de 2020. Accedido: 29 de junio de 2022. Disponible en: http://arxiv.org/abs/1912.07274

[10] D. Suleiman, W. Etaiwi, y A. Awajan, Recurrent Neural Network Techniques: Emphasis on Use in Neural Machine Translation, *Informatica*, vol. 45, n.º 7, Art. n.º 7, dic. 2021, Accedido: 29 de julio de 2022. Disponible en: https://www.informatica.si/index.php/informatica/article/view/3743

[11] N. Perez *et al.*, Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English, *Bioinformatics*, vol. 36, n.º 6, pp. 1872-1880, mar. 2020, doi: 10.1093/bioinformatics/btz853.

[12] S. Agrawal y M. Carpuat, Controlling Text Complexity in Neural Machine Translation. arXiv, 3 de noviembre de 2019. Accedido: 29 de julio de 2022. Disponible en: http://arxiv.org/abs/1911.00835

[13] L. Yeganova *et al.*, Findings of the WMT 2021 Biomedical Translation Shared Task: Summaries of Animal Experiments as New Test Set, Punta Cana, Dominican Republic, nov. 2021. Accedido: 29 de julio de 2022. Disponible en: https://hal.archives-ouvertes.fr/hal-03435096

[14] K. Bhowmik y A. Ralescu, Leveraging Vector Space Similarity for Learning Cross-Lingual Word Embeddings: A Systematic Review, Digital, vol. 1, n.o 3, Art. n.o 3, sep. 2021, doi: 10.3390/digital1030011.

[15] M. Pikuliak, M. Šimko, y M. Bieliková, Cross-lingual learning for text processing: A survey, Expert Syst. Appl., vol. 165, p. 113765, mar. 2021, doi: 10.1016/j.eswa.2020.113765.

[16] S. Pal, S. Naskar, y S. Bandyopadhyay, A Hybrid Word Alignment Model for Phrase-Based Statistical Machine Translation, en Proceedings of the Second Workshop on Hybrid Approaches to Translation, Sofia, Bulgaria, ago. 2013, pp. 94-101. Accedido: 10 de julio de 2022. Disponible en: https://aclanthology.org/W13-2814

[17] O. Franz Josef y N. Hermann, A Systematic Comparison of Various Statistical Alignment Models, 03/23. http://www.fjoch.com/giza-training-of-statistical-translation-models.html (accedido 10 de julio de 2022).

[18] P. Liang, B. Taskar, y D. Klein, Alignment by Agreement, Dep. Pap. CIS, jun. 2006. Disponible en: https://repository.upenn.edu/cis_papers/533

[19] M. Ott et al., fairseq: A Fast, Extensible Toolkit for Sequence Modeling. arXiv, 1 de abril de 2019. Accedido: 28 de junio de 2022. Disponible en: http://arxiv.org/abs/1904.01038

[20] F. Soares, V. Moreira, y K. Becker, A Large Parallel Corpus of Full-Text Scientific Articles, presentado en LREC 2018, Miyazaki, Japan, may 2018. Accedido: 28 de junio de 2022. Disponible en: https://aclanthology.org/L18-1546

[21] EMEA. https://opus.nlpl.eu/EMEA.php (accedido 14 de junio de 2022).

[22] API de Cloud Healthcare | Cloud Healthcare API, Google Cloud. https://cloud.google.com/healthcare-api?hl=es-419 (accedido 16 de julio de 2022).
[23] MetaMap. https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html (accedido 16 de julio de 2022).

[24] A. Roberts et al. The CLEF Corpus: Semantic Annotation of Clinical Text, AMIA. Annu. Symp. Proc., vol. 2007, pp. 625-629, 2007