A social media NMT engine for a low-resource language combination

María do Campo Bayón
Universitat Autònoma de Barcelona
Spain
maria.docampo@autonoma.cat

Pilar Sánchez-Gijón
Universitat Autònoma de Barcelona
Spain
pilar.sanchez.gijon@uab.cat

Abstract

The aim of this article is to present a new Neural Machine Translation (NMT) from Spanish into Galician for the social media domain that was trained with a Twitter corpus. Our main goal is to outline the methods used to build the corpus and the steps taken to train the engine in a low-resource language context. We evaluated the engine performance both with regular automatic metrics and with a new methodology based on the non-inferiority process and contrasted this information with a human evaluation based on an error classification conducted by professional linguists. We will present the steps carried out following the conclusions of a previous pilot study, describe the new process followed, analyze the new engine and present the final conclusions.

1 Introduction

In recent years, the low-resource languages domain has received some attention from our research community. Many papers covered different strategies to overcome the need for data to train engines for low-resource languages. Ranathunga *et al.* (2023) gave a complete overview of the main techniques and solutions employed in this field: data augmentation techniques, such as word or phrase replacement, back-translation, parallel data mining; unsupervised NMT; semi-supervised NMT; multilingual NMT; transfer learning in NMT; and zero-shot NMT.

A considerable amount of work has been also done in social media research, mainly in sentiment analysis and translation of user-generated content fields. The majority of these papers are focused on Phrase-Based Machine Translation (PBMT) engines. Only Lohar *et al*. (2019) attempted to compare the machine translations of tweets using phrase-based and neural MT and the usage of different amounts and types of training corpora for each of the two approaches. The results of their research showed that using a tiny Twitter corpus is useless for NMT training, although the system improved when using back-translation and out-of-domain corpora. This particular procedure is the one used in our NMT training and adapted to a low-resource language combination, from Spanish to Galician.

2 Background

This contribution presents the most significant findings from a doctoral thesis on low-resource languages and NMT as a means of promoting and using a minority language in the context of social media[1]. It is carried out in the DespiteMT project framework, dedicated to researching the uses of MT applied to the media[2]. This study is based on a previous pilot completed in 2022, which focused on creating a Spanish into Galician NMT engine for social media and proposing a new methodology for evaluating this type of NMT engine (do Campo *et al*., 2022).

For the pilot study, we created an NMT engine based on Joey through the online platform Mut-NMT[3] (a minimalist NMT toolkit for novices, https://aclanthology.org/D19-3019/). The corpus used to train the engine was a mix of two corpora. We used the Paracrawl corpus Spanish – Galician (1,879,651 sentences and 44,626,394 words) as a generic base corpus. To build a social media corpus,

[1] Ph.D. Title: *NMT for low-resource languages. Optimization and strategies*. Universitat Autònoma de Barcelona. Forthcoming.

[2] DespiteMT project, grant number PID2019-108650RB-I00 [MINECO / FEDER, UE; Principal researcher: Dr. Pilar Sánchez-Gijón, Grup Tradumàtica, UAB.

[3] Available at: https://www.multitrainmt.eu/index.php/es/formacion-en-ta-neuronal/mutnmt and explained in Kenny, 2022

we decided to extract Galician tweets from Twitter as finding parallel corpora would be very difficult. The idea was to find monolingual in-domain text and then back-translate it. Thus, we first created a Galician monolingual corpus of tweets written in Galician and extracted from Galician six institutional accounts (see **Table *1***): three accounts from the Galician Government and linguistic institutions and the accounts from the three Galician universities.

Table 1: Number of tweets crawled per Galician institutional account

Twitter account	Number of tweets
@uvigo	11507
@UDC_gal	10258
@UniversidadeUSC	7875
@AcademiaGalega	6362
@PortalPalabras	5543
@SXPL	4165

These accounts were chosen specifically as they would be more reliable in terms of good use of grammar and spelling, as well as common and natural expressions. To extract the text of tweets and hashtags, we used the Python library *snscrape*, which is a scraper for social networking services (SNS) and scrapes things like user profiles, hashtags, or searches and returns the discovered items, e.g. the relevant posts. It allowed us to specifically target the desired accounts and crawled all tweets of their accounts. The resulting file was a JSON file that we converted into a CSV file to handle the text. In the CSV file, we also erased all content except the tweet and the hashtags, eliminated URLs and icons, and finally checked that there was no content in other languages.

After cleaning the monolingual corpus, we back-translated it into Spanish using the generic Google Translate engine to give us a bilingual corpus (69,713 unique sentences). Then, we use the MTUOC python library[4] to process the bilingual corpora and prepare it to train our engine. The engine is available at: https://ntradumatica.uab.cat/.

Once trained, the engine achieved a total BLEU (Papineni *et al*., 2002) score of 70.63 against the test corpus extracted only from the Twitter corpus. We also conducted an end-user evaluation based on the non-inferiority principle (do Campo *et al*., 2022). In pharmaceutical studies, this is commonly used to determine whether a treatment or product is not worse than an active treatment or product. The pilot study had two main objectives: validate the method and evaluate our NMT engine. In this study, the non-inferiority principle attempted to determine whether tweets generated by NMT are perceived as inferior (Molina, 2020; Althunian *et al*., 2017; Tunes da Silva *et al*., 2009) or less natural than tweets directly written in Galician. From a pragmatic point of view, non-inferiority stands for MT-obtained texts which are not perceived are less natural than any other piece of text originally written in the target language. The sample of tweets was selected following two criteria. First, they were classified according to their origin: original text if the text was directly written in Galician, and machine translation if the text was machine translated from Spanish into Galician using our NMT. Then, they were classified according to their length: short sentence, long sentence, paragraph composed of short sentences, paragraph composed of long sentences, and mixed paragraph if the paragraph contained both short and long sentences.

According to the results of this previous pilot study, we were able to draw several conclusions. On the one hand, we found weaknesses and strengths of the performance of the NMT engine in a low-resource language context. The estimations based on the model (do Campo *et al.*, forthcoming) indicated the path to improving our engine. The performance in short sentences presented both individually or in a paragraph should be improved in order to reach non-inferiority in all kinds of tweets. Surprisingly, we discovered that our engine was not inferior to tweets directly written in Galician and formed by long sentences. On the other hand, we validated our analysis method. We demonstrated that non-inferiority evaluations can be used to extract end-user perceptions in machine translation evaluation.

Hence, we designed the final training and repeated our study taking into account the pilot conclusions.

3 Retraining process of the NMT engine for social media

In the second training of the NMT engine, we changed two settings of the first NMT engine setup: the amount of data of the specific corpus and the NMT technology. We kept the Paracrawl corpus as a base generic corpus but decided to expand our Twitter corpus. Our first Twitter corpus contained only nearly 70000 unique sentences. To build a larger Twitter corpus, we chose more institutional Galician accounts (see **Table *2***), such as those associated with Galicia's official television and radio, accounts designed to promote Galician, divulgation magazine accounts, and podcasts in Galician accounts. These 18 accounts were

[4] Available here: https://github.com/aoliverg/MTUOC.

specifically chosen following the same criteria as in the previous study of good use of grammar and spelling, as well as common and natural expressions.

Table 2: Number of tweets crawled per Galician institutional account

Twitter account	Number of tweets
@uvigo	11507
@UDC_gal	10258
@UniversidadeUSC	7875
@AcademiaGalega	6362
@PortalPalabras	5543
@SXPL	4165
@Falaredes	8887
@culturagaleg	28252
@consellocultura	5486
@biosbardia	4191
@EdGalaxia	13518
@podgalego	5644
@comochodigo	362
@ctnl	9994
@NeoFalantes	362
@GalegoTwitch	13223
@diariocultural_	13545
@RadioGalega	81604
@TVGalicia	144741
@DigochoEuTVG	738
@IGE_Estatistica	27131
@Valedordopobo	3385
@Fegamp	3743
@Par_Gal	17258

They also are very active accounts with much more content. We used the same Python scripts to crawl the content and clean the resulting corpus (see **Figure *1***). First, we manually crawled the specific accounts, exported the content into JSON files, and then converted them into CSV.

Snscrape.py

```
import os
import pandas as pd
tweet_count = 100000
username = "Par_Gal"
os.system("snscrape --jsonl --max-results 100000 twitter-search
'from:Par_Gal'> text19.json".format(tweet_count, username))

# Reads the json generated from the CLI command above and creates a pandas
dataframe
tweets_df1 = pd.read_json('text19.json', lines=True)
tweets_df1.to_csv('text19.csv', sep=',', index=False)
```

Figure 1: Example of the Python script used to crawl the tweets of the Galician Parliament account

Second, handled the CSV content (see **Figure *2***) and eliminated everything except the content of the tweet. Then, we cleaned URLs and icons, and checked that all tweets were in fact written in Galician, as we found some content in Spanish in the previous crawling.

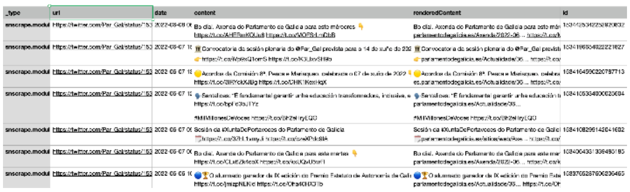

Figure 2: CSV file created from the JSON export

After erasing URLs, icons, and other language content, we back-translated the tweets into Spanish with Google Translate. We used this technique because of the good results obtained in the first evaluation and in the bibliography. We obtained a bilingual file of 299,051 translation units. To clean and tokenize the bilingual corpus, we used the MTUOC library. The MTUOC clean script allowed us to normalize apostrophes, remove HTML/XML tags, unescape html entities and remove segments with empty source or target. It also allowed us to remove source and target segments that were equal. The cleaned bilingual corpus contained 262,785 unique sentences and 5,448,375 words. We trained our engine with the generic Paracrawl corpus Spanish – Galician (1,879,651 sentences and 44,626,394 words) and this specific corpus.

Regarding the NMT technology, we used a transformer-big configuration for Marian and sentencepiece (Wolf *et al*., 2020). We decided to change the NMT technology used to have better control of the training parameters as the Joey MutNMT platform does not allow this. The BLEU score obtained was 85% against the test corpus extracted only from the Twitter corpus, which was higher than the BLEU score achieved in the pilot study.

4 NMT engine evaluation

We carried out two different types of evaluation of our NMT engine. First, we replicated the non-inferiority evaluation presented in our first study with some adjustments and obtained better results (do Campo *et al.*, forthcoming). Second, we conducted an error evaluation using the DQF-MQM framework (Lommel *et al.*, 2018; Görög, 2014; Popovic, 2018), which was carried out by three professional linguists with over ten years of experience using the online platform ContenQuo. ContentQuo is an online platform, specially dedicated to translation quality evaluation with specific workflows and predefined quality templates, such as the one used, DQF. The goal of doing two different evaluations was to find a link between errors and a negative attitude toward the machine-translated tweets, similar to the methodology applied by Guerberof *et al.*, 2022 and Bhardwaj *et al.*, 2020.

To carry out the error-classification evaluation based on the DQF-MQM framework, we selected three Galician linguists with more than 10 years of experience in the Spanish-Galician language direction and with experience in Machine Translation through Proz.com. The professional evaluation was remunerated and conducted in the app ContentQuo[5] (see **Figure *3***).

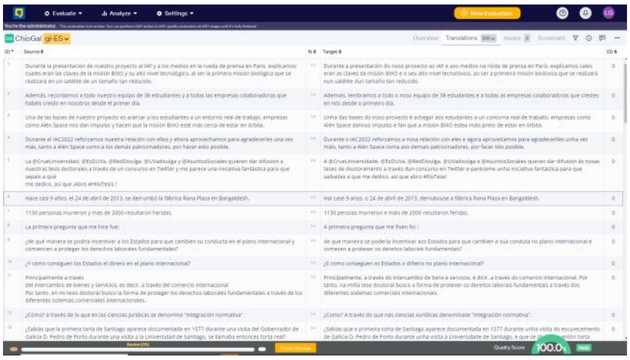

Figure 3: ContentQuo interface

We asked proofreaders to review 30 tweets that were translated from Spanish into Galician using our NMT engine for social media (890 words). Linguists were asked to assess the raw MT output. Those tweets were the same used in the non-inferiority evaluation survey. A brief explanation before the evaluation was given to contextualize the task and explain the objective of the assignment. They dedicated one to two hours to this task.

The mean overall quality score obtained was 94.55%. Although this is a good score taking into account that the tweets were not postedited, we were more interested in the type of errors annotated by the professional linguists and in the severity of the errors (see **Table *3***). No errors were found in the following DQF-MQM categories: verity, locale convention, and design. Some errors were found in the categories style and others (errors that cannot be categorized in any of the rest of the categories). As expected, most of the errors were found in the fluency, adequacy, and terminology categories. Grammar, punctuation, and spelling errors were found in the fluency category, while mistranslations and over-translations were found in the adequacy category.

Regarding the severity, no critical errors were found and the majority of the errors were minor. Only a few errors were classified as major. The DQF-MQM template also allowed the classification of the errors as neutral without affecting the quality score. In **Table *3***, a detailed list of errors by error category is presented.

Table 3: list of errors found in the error classification evaluation using the DQF-MQM framework

Error Category	Neutral	Minor	Major	Critical
Verity	0	0	0	0
Terminology	0	2	2	0
Style	1	2	0	0
Other	3	3	0	0
Locale convention	0	0	0	0
Fluency	0	14	2	0
Design	0	0	0	0
Accuracy	2	4	3	0

An intriguing finding was that major errors were mostly distributed in threads and short paragraphs, which could explain the survey's low acceptance (do Campo *et al.*, forthcoming). Furthermore, while minor errors were distributed indiscriminately in all types of tweets, it appears that this severity of errors has no effect on users' perceptions of naturalness.

We also asked the linguists for a general comment about the quality of the raw machine-translated tweets. Generally speaking, they agreed on the good quality of the engine. They highlighted that some segments did not need any change, while others need a few changes to be correct with respect to the Spanish text.

5 Conclusions

The purpose of this article was to describe the process of developing an NMT engine for social media in a low-resource language combination using Twitter data and back-translation as primary strategies. We have shown that increasing the in-domain Twitter corpus and using back-translation improved the

[5] Available here: https://www.contentquo.com/

engine's performance in terms of both automatic and human evaluation. We also want to emphasize that the size of the in-domain Twitter corpus will be determined by the proximity of the languages used. As Spanish and Galician are very close languages, we saw promising results with a small in-domain corpus.

As shown by us and other authors, Twitter is a good source of monolingual data crawling. In this article, we have shown that it can be used for more than common uses such as sentiment analysis.

Furthermore, professional linguists concluded that the raw machine-translated tweets evaluated could benefit from minor post-editing. The double evaluation conducted −non-inferiority (do Campo et al., forthcoming) and human evaluation− demonstrated that our engine is capable of translating social media content.

Finally, we want to contextualize the importance of conducting NMT research on low-resource languages in order to promote their use. Both our training process and evaluation methodology can be replicated in other language combinations that are similar to ours, particularly if they want to promote the low-resource language on social media and the Internet. With our research, we hope to help Galician to reach the younger population through social media and reduce the loss of speakers in the last decades.

References

Althunian, Turki A., de Boer, Anthonius, Groenwold, Rolf and Olaf H. Klungel. 2017. Defining the noninferiority margin and analysing noninferiority: an overview. In *British Journal of Clinical Pharmacology*, 83(8): 1636-1642.

Bhardwaj, Shivendra, Hermelo, David A., Langlais, Phillippe, Bernier-Colborne, Gabriel, Goutte, Cyril and, Michel Simard. 2020. Human or Neural Translation? In *Proceedings of the 28th International Conference on Computational Linguistic*s. 6553–6564 Barcelona, Spain.

Chatzikoumi, Eirini. 2020. How to evaluate machine translation: A review of automated and human metrics. In *Natural Language Engineering*, 26(2): 137–161. Cambridge University Press.

do Campo Bayón, María, and Pilar Sánchez-Gijón. 2022. Evaluating NMT: Superior, Inferior, or Equivalent to Texts Originally Written by Humans. In *Proceedings of the New Trends in Translation and Technology, NeTTT 2022*. Rhodes, 4-6 July.

do Campo Bayón, María. and Pilar Sánchez-Gijón. Evaluating NMT using the non-inferiority principle (forthcoming).

Guerberof-Arenas, Ana. and Antonio Toral. 2022. Creativity in translation: machine translation as a constraint for literary texts. In *Translation Spaces*. John Benjamins Publishing Company.

Hassan, Hany, Aue Anthony, Chen, Chang, Chowdhary, Vishal, Clark, Jonathan, Federmann, Christian, Huang Xuedong, Junczys-Dowmunt, Marcin, Lewis, William, Li, Mu, Liu, Shujie, Liu, Tie-Yan, Luo, Renqian, Menezes, Arul, Qin, Tao, Seide, Frank, Tan, Xu, Tian, Fei, Wu, Lijun, Wu, Shuangzhi, Xia, Yingce, Zhang, Dongdong, Zhang, Zhirui, and Ming Zhou. 2018. *Achieving human parity on automatic Chinese to English news translation*.

Görög, Attila. Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. In *Tradumàtica* 12 (2014): 443-454.

Kenny, Dorothy. 2022. *Machine translation for everyone: Empowering users in the age of artificial intelligence*. (Translation and Multilingual Natural Language Processing 18). Berlin: Language Science Press.

Lohar, Pintu, Popović, Maja, Alfi, Haithem, and Andy Way. 2019. A systematic comparison between SMT and NMT on translating user-generated content. In 2*0th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, 7 - 13 Apr 2019, La Rochelle, France.

Lommel, Arle, and Alan Melby. 2018. Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas* (Volume 2: User Track). Boston.

Maruf, Sameen, Saleh, Fahimeh, and Gholamreza Haffari. 2021. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. In *ACM Computer. Survey*. 54, 2, Article 45 (March 2022), 36 pages.

Molina Nadal, Andrea. 2020. Ensayos clínicos de no inferioridad. In *FMC - Formación Médica Continuada en Atención Primaria*, 27(7): 345-348.

Papineni, Kishore, Roukos, Salim, Ward, Todd, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation ofmachine translation. In P*roceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318.

Popović, Maja. 2018. Error classification and analysis for machine translation quality assessment. In *Translation quality assessment, Machine Translation: Technologies and Applications* (1): 129– 158. Springer, Charm.

Ranathunga, Surangik, Lee, En-Shiun Annie, Skenduli, Marjana Prifti, Shekhar, Ravi, Alam, Mehreen, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-Resource Languages: A Survey. In *arXiv e-prints*.

Rivera-Trigueros, Irene. 2021. Machine translation systems and quality assessment: a systematic review. In *Language Resources and Evaluation*, 1–27.

Tan, Xin, Zhang, Longyin, and Guodong Zhou. 2022. Discourse Cohesion Evaluation for Document-Level Neural Machine Translation. In *arXiv preprint arXiv*: 2208.09118.

Toral, Antonio, Castilho, Sheila, Hu, Ke, and Andy Way. 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, Volume 1: 113-123. Research Papers, Association for Computational Linguistics, Brussels, Belgium.

Tunes da Silva, Gisela, Logan, Brent, and John P Klein. 2009. Methods for equivalence and noninferiority testing. In *Biol Blood Marrow Transplant*, 15(1 Suppl): 120-127.

Van der Werff, Tobias, van Noord, Rik, and Antonio Toral. 2022. Automatic Discrimination of Human and Neural Machine Translation: A Study with Multiple Pre-Trained Models and Longer Context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 161-170, Ghent, Belgium. European Association for Machine Translation.

Van Edgom, Gys-Walt and Mark Pluymaekers. 2019. Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end users. In *The Journal of Specialised Translation*, 31: 158-176.

Way, Andy. 2018. Quality expectations of machine translation. In *Translation quality assessment, Machine Translation: Technologies and Applications* (1): 159–178. Springer, Charm.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pier- ric Cistac, Tim Rault, Rémi Louf, Morgan Funtow- icz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language pro- cessing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.