Is it safe to machine translate suicide-related language from English to Galician?

John E. Ortega
Northeastern University
Boston, MA, USA
j.ortega@northeastern.edu

Annika Marie Schoene
Northeastern University
Boston, MA, USA
a.schoene@northeastern.edu

Abstract

In this article, we present work that uses a pre-trained language model (PLM) from one of the most widely-used machine translation (MT) systems to translate suicide-related language from an English lexicon to Galician, a language commonly spoken in northern Spain. We make the MT system translations publicly available along with other annotations from professional Galician translators. Additionally, we compare and contrast the findings to provide insight into the types of errors that a MT system may commit when translating from English to Galician in life-threatening situations.

1 Introduction

With the widespread use of large and pre-trained language models (LLMs and PLMs) it is often the case that the assumption: "something is better than nothing" for machine translating low- to medium-resource languages is a safe assumption. The assumption relies on the fact that most low-resource models do not perform well under the constraint of scarce resources. Hence, projects like the *No Language Left Behind* (NLLB) project (Costa-jussà et al., 2022) offer models with billions of parameters that are trained on large amounts of monolingual data in self-supervised learning (SSL) manner to give *some* information that in turn will produce better translations into the low-resource target language then a sustained parallel model created from a few thousand parallel (low-resource to high-resource) translations.

While it may be the case that for many generic situations, minimal translations are better than none at all, we argue in this article that for crisis situations where a human life can be at stake, it may be better to first consider the quality of the output from a machine translation (MT) system, for example. In order to better support our argument, we create translations from English to Galician, a low-to-medium resource language spoken mostly in the north of Spain, by using a state-of-the-art MT system based on a PLM. We then verify the validity of the output in an annotation task where we ask native Galician translators to provide honest feedback concerning several key metrics often used in MT. Finally, we present our experimental results in a public corpus which provides the English lexicon, its Galician translations, and the final feedback available online.

To this end, we first present unprecedented work presented by others in Section 2. We feel that it is important to review the work in Section 2 to get an idea of how much need is warranted for corpora similar to the one we present in this article. The corpus creation and collection details along with the MT system used are then presented in Section 3. Finally, in Section 4, we provide insight into the findings of how LLMs/PLMs performed in our experiments and then conclude our work in Section 5.

2 Related Work

Since mental health and suicide can be found to be a challenging and even "modern" topic, the amount of literature currently available is somewhat dearth. Nonetheless, we present work that is directly related to the region (Spain), where Galician is spoken rather than reporting on English suicide-related work. For further exploration, we provide the lexicon in Appendix A.

The broader topic that our works is related to is called: *suicide ideation* (SI). Wikipedia defines SI as "suicidal thoughts or the thought process of having ideas, or ruminations about the possibility of ending one's own life".[1] Other suicide professionals (Klonsky et al., 2016) including the Center for Disease Control first define suicide as "death

[1] https://en.wikipedia.org/wiki/Suicidal_ideation

caused by self-directed injurious behavior with an intent to die as a result of the behavior" and then SI as "thinking about, considering, or planning suicide". In SI, recent work has been performed on suicidal notes in Spanish (Valeriano et al., 2020; Ramírez-Cifuentes et al., 2020) that led to novel findings for English–Spanish classification. Their work did not cover Galician but can serve as a broader baseline for machine learning approaches that could use the corpora we present in Galician. While somewhat less recent, Fernández-Cabana et al. (2015) translated Galician to Spanish suicide notes due to the lack of parallel resources available. It is our opinion, that this provides motivation to create the corpora and annotations like those presented here. The lack of resources available in Galician does not coincide with the need. Recent research by Flórez et al. (2023) has shown that in Spanish autonomous communities like Galicia, there are high suicide rates on the order of 12 to 13 percent per 100,000 inhabitants. We are not sure, but the suicide rates may somehow relate to the lack of digital and economic resources as reported by other work (Fernández-Navarro et al., 2016).

Given the scarcity of the work directly related from English to Galician where English generally has more resources, we feel that this article is the first in a series of publications that direct its efforts to investigating mental health issues in low-to-medium resource communities like Galicia.

3 Methodology

In this section, we describe the steps taken to reproduce our work which first begins with the lexicon and MT system used. Secondly, we demonstrate how we evaluated the output from the MT system with native Galician translators and lastly we discuss the metrics we used to compare the validity of the translations and human opinions.

3.1 Lexicon and MT System

We detail the process used for first creating the Galician texts based on the original lexicon in English. The original lexicon can be found in Appendix A. It contains 50 phrases related to suicidal ideation and was proposed by O'dea et al. (2015).

In order to better evaluate how MT performs with a PLM on suicide-related phrases, we use widely-used translation toolkit from Facebook called *Fairseq*[2] (Ott et al., 2019). More specifically, we use the default settings proposed by the NLLB research group (Team et al., 2022). To our knowledge, the Fairseq/NLLB MT system is the best-performing system for low-resource languages, including some medium-resource languages like Galician. It has a transformer-based (Vaswani et al., 2017) PLM and would be easy-to-use quickly in crisis situations. While we do not necessarily recommend its use in crisis situations, we recognize that the tool would more than likely be the first one used by MT researchers in the field in a crisis situation (with the exception of a few others that have not released resources in an open-source manner).

3.2 Evaluation

We conducted a round of pilot evaluations of our proposed qualitative measures by inviting native Galician translators to perform manual evaluations of the translated lexicon. Annotators are given (i) the original dictionary, (ii) the translated dictionary and (iii) a codebook with an example evaluation for reference[3]. At this stage, we specifically did not ask for translators with experience medical, psychological or behavioral health training experience. This is due to the short phrases and background of terms in the original lexicon being everyday language albeit suicide-related language. In total, there were two annotators. For each dictionary entry we asked annotators to consider five variables that focus on the following aspects:

- **Adequacy** Similar to Castilho et al. (2018) we asked annotators to rate on a Likert scale how adequate a translation is by asking *'How much of the source text meaning has been retained in the translated language?'*.
- **Fluency** Annotators were asked how fluent translations were and asked to rate them on a Likert scale.
- **Spelling Errors** When translations contained errors, such as 'misspelled words', 'missing words', 'added words' or 'incorrect word order', annotators were asked to score from 0 to 1.
- **Cultural Acceptability** Since suicidal language is not universal (Kirtley et al., 2022) and depends on cultural context, we asked the annotators to provide a "yes" only when the

[2] https://github.com/facebookresearch/fairseq

[3] https://github.com/annikamarie/MultiLingual-SI

Original word/phrase	Proposed Translation	Alternative Translation
to take my own life	para quitar a miña propia vida	para quitarme a vida
slit my wrist	cortoume o pulso	cortar o pulso
go to sleep forever	Vai durmir para sempre	vai adormecer para sempre

Table 1: Examples of alternative translations contributed by annotators.

source language's intent matched the target language's intent.

- **Context** Along with cultural appropriateness, we asked the annotators to verify that the translated context captures the suicide-related language with a "yes" or "no". Additionally, annotators were given the option to add a *new* variation of the original lexicon if they saw it was necessary.
- **Alternative Translation** Annotators were asked to provide feedback and alternative translations when saw fit. We collected their comments and translations.
- **Contributions in local language** Annotators were asked to add (i) words related to death, suicide, and/or (ii) expressions/metaphors related to dying and (iii) expressions/metaphors related to suicide.

In order to better illustrate the alternative contributions from annotators, an example is provided in Table 1 of a few words taken from the original lexicon along with their translation and alternative translation provided by an annotator.

3.3 Metrics

In order to measure the performance of the Fairseq/NLLB system output when compared to the annotator's feedback, we used a constant metric over all of the items annotated. We represent the total number of entries from the original lexicon as N. For the submitted evaluations E, we calculate the arithmetic mean $\overline{x}$ as follows:

$$\overline{x} = \frac{\sum_E}{N} \quad (1)$$

The scores for each annotator attribute are:

- **Adequacy** and **Fluency** – A dictionary can score a maximum value of *4*, meaning all meaning and fluency has been retained in the translation respectively.
- **Spelling Errors** – When no spelling errors are made, a score of *0* is recorded.
- **Cultural** and **Contextual acceptability** – When all translations are deemed appropriate, the best score is *1*.

4 Results

In this section we present the initial results achieved by our pilot study. While the results presented here do not constitute an indication of the Fairseq/NLLB performance being better for Galician in a crisis situation, we feel that our study has shown that for a small set of translated phrases, its initial performance can be considered *helpful*. At this point, we leave further investigation which would involve clinicians and other more qualified professionals as future work.

The results in Table 2 provide the evidence from our pilot study. *Adequacy* and *fluency* score well since the maximum for both is considered to be a score of 4. While some spelling errors were prevalent, we are happy to report that the cultural context seems to have been captured (something we did not initially expect). One additional result not reported in Table 2 is that we received alternative translations for 23 terms. These will be added to the translated lexicon and shared publicly for others to use. We provide three examples of those alternatives in Table 1.

Variable	Result
Adequacy	3.48
Fluency	3.41
Spelling Errors	0.29
Culture	0.96
Context	0.96

Table 2: Evaluation of translated Galician dictionary using quantitative metrics.

5 Conclusion

In this paper, we have shown that for at least one crisis situation with suicide-related language, MT may be more likely used and can provide some initial insight into the type of language being used. To the same end, we have also shown that for even a small list of phrases, MT system translations are not perfect. Thus, it would be better to have a

human in the loop if possible to correct or suggest more translations.

We provide our dataset and annotations publicly and invite future investigations to use it. It is our hope to create a larger consortium to collaborate at an international level to help better understand suicide ideation and other factors related to it. In future work, we will develop a more comprehensive set of ethical guidelines, improve the lexicon quality of other languages, and compare other MT systems to the Fairseq/NLLB one. Additionally, it is our aim to get help from clinical professionals trained in languages like Galician to participate in further studies.

Ethical Considerations There are many considerations when engaging with automated multilingual suicide ideation detection, which can relate but are not limited to (i) concerns related to linguistic aspects (e.g.: linguistic imbalances and misrepresentation) and (ii) concerns related to developing, designing, and deploying dictionaries to the public (e.g.: issues of autonomy, justice and harms), especially given their usefulness to build automated tools for suicide detection.

References

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou. 2018. Evaluating mt for massive open online courses: A multifaceted comparison between pbsmt and nmt systems. *Machine translation*, 32(3):255–278.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mercedes Fernández-Cabana, Julio Jiménez-Féliz, María Teresa Alves-Pérez, Raimundo Mateos, Ignacio Gómez-Reino Rodríguez, and Alejandro García-Caballero. 2015. Linguistic analysis of suicide notes in spain. *The European Journal of Psychiatry*, 29(2):145–155.

P Fernández-Navarro, ML Barrigón, J Lopez-Castroman, M Sanchez-Alonso, M Páramo, M Serrano, M Arrojo, and E Baca-García. 2016. Suicide mortality trends in galicia, spain and their relationship with economic indicators. *Epidemiology and psychiatric sciences*, 25(5):475–484.

Gerardo Flórez, Ashkan Espandian, Noelia Llorens, Teresa Seoane-Pillado, and Pilar A Saiz. 2023. Suicide deaths and substance use in the galician provinces between 2006 and 2020. *Frontiers in psychiatry*, 14:1242069.

Olivia J Kirtley, Kasper van Mens, Mark Hoogendoorn, Navneet Kapur, and Derek de Beurs. 2022. Translating promise into practice: a review of machine learning in suicide research and prevention. *The Lancet Psychiatry*, 9(3):243–252.

E David Klonsky, Alexis M May, and Boaz Y Saffer. 2016. Suicide, suicide attempts, and suicidal ideation. *Annual review of clinical psychology*, 12:307–330.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Diana Ramírez-Cifuentes, Ana Freire, Ricardo Baeza-Yates, Joaquim Puntí, Pilar Medina-Bravo, Diego Alejandro Velazquez, Josep Maria Gonfaus, and Jordi Gonzàlez. 2020. Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. *Journal of medical internet research*, 22(7):e17758.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kid Valeriano, Alexia Condori-Larico, and Josè Sulla-Torres. 2020. Detection of suicidal intent in spanish language social networks using machine learning. *International Journal of Advanced Computer Science and Applications*, 11(4).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Example Appendix

- suicidal, kill myself, my suicide letter, end my life, never wake up, suicide pact, die alone, wanna die, why should I continue living, to take my own life, suicide, can't go on, want

to die, be dead, better off without me, better off dead, dont want to be here, go to sleep forever, wanna suicide, take my own life, suicide ideation, not worth living, ready to jump, sleep forever, suicide plan, tired of living, die now, commit suicide, thoughts of suicide, depressed, slit my wrist, cut my wrist, slash my wrist, do not want to be here, want it to be over, want to be dead, nothing to live for, ready to die, not worth living, I wish I were dead, kill me now, hit life, think suicide, wanting to die, suicide times, last day, feel pain point, alternate life, time to go, beautiful suicide, hate life