Proxecto Nós: Artificial intelligence at the service of the Galician language

Proxecto Nós: Inteligencia artificial al servicio de la lengua gallega

Adina Ioana Vladu [1], Iria de-Dios-Flores [2], Carmen Magariños [1], John E. Ortega [2], José Ramom Pichel [2], Marcos Garcia [2], Pablo Gamallo [2], Elisa Fernández Rei [1], Alberto Bugarín [2], Manuel González González [1], Senén Barro [2] and Xosé Luis Regueira [1]

[1] *Instituto da Lingua Galega (ILG) - Universidade de Santiago de Compostela, Spain*
[2] *Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) - Universidade de Santiago de Compostela, Spain*

Abstract

Proxecto Nós is an initiative aimed at providing the Galician language with openly licensed resources, tools, and demonstrators in the area of intelligent technologies. The Project has two main scientific and technological objectives: (i) to integrate the Galician language into cutting-edge AI and language technologies, thus enabling the natural use of Galician in human-machine interactions; and (ii) to improve the state of the art of language technologies for Galician.

Keywords

Language technologies, linguistic rights, Galician, low-resource languages.

1. Introduction

Proxecto Nós (The *Nós* Project) is an initiative promoted by the Galician Government (Xunta de Galicia), aimed at providing the Galician language with openly licensed resources, tools, demonstrators, and use cases in the area of intelligent technologies. The execution of *Proxecto Nós* has been entrusted to the University of Santiago de Compostela (USC) and is currently being carried out by a research team comprising members of the Instituto da Lingua Galega (ILG) and the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS). The first stage, spanning from the final trimester of 2021 to 2025, will lay the foundations and provide the resources that will help place Galician among the languages that are fully active in the digital society and economy.

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
EMAIL: adina.vladu@usc.gal (A.I. Vladu); iria.dedios@usc.gal (I. de-Dios-Flores); mariadelcarmen.magarinos@usc.gal (C. Magariños); john.ortega@usc.gal (J. Ortega); jramom.pichel@usc.gal (J.R. Pichel); marcos.garcia.gonzalez@usc.gal (M. Garcia); pablo.gamallo@usc.gal (P. Gamallo); elisa.fernandez@usc.gal (E. Férnandez Rei); alberto.bugarin.diz@usc.gal (A. Bugarín); manuel.gonzalez.gonzalez@usc.gal (M. González González) senen.barro@usc.gal (S. Barro); xoseluis.regueira@usc.gal (X.L. Regueira)
ORCID: 0000-0002-3910-7820 (A.I. Vladu); 0000-0002-5941-1707 (I. de-Dios-Flores); 0000-0003-3525-1304 (C. Magariños); 0000-0002-2328-3205 (J. Ortega); 0000-0001-5172-6803 (J.R. Pichel); 0000-0002-6557-0210 (M. Garcia); 0000-0002-5819-2469 (P. Gamallo); 0000-0002-4109-0087 (E. Férnandez Rei); 0000-0003-3574-3843 (A. Bugarín); 0000-0001-7948-4607 (M. González González) 0000-0001-6035-540X (S. Barro); 0000-0001-7264-3740 (X.L. Regueira)

2. Context and motivation

The development of language technologies is a strategic innovation area geared towards the digital society and economy, and it has been a priority in both Spanish (Plan Estatal de Investigación Científica y Técnica y de Innovación, Estrategia Española de Ciencia y Tecnología y de Innovación) and European (Horizon 2020) scientific planning. Technologies such as machine translation (MT), information extraction (IE), text analytics, and dialogue systems are essential in the digital society, culture, and economy.

Languages in high demand worldwide (especially English) benefit from a large variety of computational resources that can contribute to developing new automatic language processing technologies and tools. Such is the case due to the long-standing research tradition in these areas (e.g., the variety of projects financed by USA's DARPA) and the need to incorporate such languages into the AI applications associated with the latest electronic devices (such as the conversational AI or automatic dictation software developed by Google, Amazon or Apple). Other languages that have joined AI research later, such as Chinese, are currently following in the footsteps of English, through projects such as Baidu's Qian Yan, which improve significantly the computational resources available in their respective language varieties.

Notwithstanding, language technologies are also necessary for languages in lower international demand. Consequently, different languages have developed similar initiatives to Nós. Among others, we can highlight Projecte AINA, which will develop computational resources for Catalan until 2024, or the work carried out at the HiTZ Research Center, focusing on languages technologies for Basque. Other projects, such as CorCenCC (in Great Britain, for Welsh) or UQAILAUT (in Canada, for Inuktitut) were considered success cases in the promotion of the digital use of socially threatened languages.

The democratization of language technologies has a great social and cultural impact on the communities that use them. For instance, MT increases access to contents in different languages, thus facilitating intercultural relations; dialogue systems allow us to communicate with machines in our own language; and semantic technologies enable advances in the automatic comprehension of texts, thus making it possible to process enormous quantities of documents. In the case of Galician, incorporating the language into state-of-the-art AI applications can not only significantly favor its prestige (a decisive factor in language normalization), but also guarantee citizens' language rights and reduce social inequality.

In economic terms, the global Natural Language Processing (NLP) market size was valued at more than USD 10 billion in 2020 and is expected to reach USD 41 billion by 2025 (Aldabe et al., 2021). NLP technologies are used in different areas such as information retrieval, MT, IE (with notable growth in its application in the medical domain during the Covid-19 pandemic), dialogue systems, and automatic text generation, among many others. The capacity to model language, an essential ability for human beings, ensures a promising future for such technologies from both an economic and research and innovation perspective.

3. State of the art: Galician resources and technologies

In 2012, the White Paper *The Galician Language in the Digital Age* (García-Mateo et al., 2012) described Galician as a language with a level of technological support that "gives rise to cautious optimism", while highlighting the need for new resources and tools. Previous research projects on Galician resulted in speech processing resources (COTOVÍA), an annotated reference corpus (CORGA), morphosyntactic lemmatizers and taggers (XIADA, FreeLing, IXA-Pipes), other specialized corpora, both text (CLUVI, CTG. TreeGal) and speech (CORILGA, AGO), MT systems (GAIO, OpenTrad), spellcheckers (OrtoGal), grammar checkers (Avalingua), language analysis and IE tools (Linguakit), language models (SemantiGal, Bertinho), and other resources.

Furthermore, Galician is currently part of multilingual crowdsourced data collection initiatives carried out by important companies on the global IT market, which have resulted in speech databases such as Google's SLR77 (Kjartansson et al., 2020) and Mozilla's CommonVoice 7.0 and 8.0 (Ardila et al., 2020). This situation is reflected in a recent report on the current state of the LT (Language Technology) field for Galician (Ramírez Sánchez & García Mateo, 2022), which informed on the considerable growth in the production of high-

quality Galician resources and services, especially text resources.

Despite the quality of these resources, it should be noted that not all are freely and publicly available for the development of LT. The LT field has undergone profound changes over the last few years since the introduction of neural network systems. Generally, training models using these state-of-the-art technologies requires large quantities of data and has high energetic and computational costs, which continues to be a challenge for low-resource languages. However, as many recent studies show, end-to-end technologies and open-source multilingual pre-trained models created using large quantities of data from high-resource languages (Shen et al., 2018; Baevski et al., 2020; Wolf et al., 2020) can be used, through transfer learning and fine-tuning, to train models in low- or medium-resource languages such as Catalan (Külebi & Öktem, 2018; Külebi et al., 2020) or, in our case, Galician. To this end, the existence of resources and tools that are freely available to the scientific and business community is essential, and that constitutes one of the main objectives of *Proxecto Nós*.

4. Project description

4.1. Organization

The tasks that are to be carried out as part of the Project can be included in the following areas, corresponding to some of the major NLP fields:

An example of numbered list is as following.

1. Speech synthesis (TTS)
2. Speech recognition (ASR)
3. Automatic text generation
4. Dialogue systems
5. MT
6. IE
7. Opinion mining and fact checking
8. Language correction and assessment

These broad, mutually interdependent areas fall within the three strategic lines jointly identified by the Project's research team and the Xunta de Galicia (in particular, with the Axencia para a Modernización Tecnolóxica de Galicia): (i) spoken or written conversation with people, (ii) language quality, and (iii) information management.

In accordance with the funding agreement signed by the Xunta de Galicia and the USC, the organization of the tasks included in *Nós* follows a yearly schedule. Each year, resources, language models and demonstrators from different areas will be made publicly available.

More information on the organization of *Proxecto Nós* can be found in de-Dios-Flores et al, 2022.

4.2. Scientific and technological objectives

Proxecto Nós has two main scientific and technological objectives: (i) to integrate the Galician language into cutting-edge AI and language technologies, thus enabling the natural use of Galician in human-machine interactions; and (ii) to improve the state of the art of language technologies for Galician.

For this purpose, resources, tools, and applications will be developed and distributed under open licenses, which will allow them to be integrated into existing devices and services (such as smart speakers or conversational agents) and future technologies. To this end, specific objectives directly related to some of the major tasks of NLP have been established.

Each of these technological objectives will be executed in a different subproject, which will allow the parallel development of different tasks and, overall, a more effective organization of the work. However, a set of general objectives are shared by all the tasks. These objectives are: (i) the compilation of high-quality linguistic resources (annotated reference corpora, web-scale corpora, specialized corpora by tasks and domains, parallel corpora, knowledge bases, dictionaries, etc.); (ii) the elaboration of language and acoustic models (both general-purpose and task-specific models); and (iii) the development of applications based on these models. The project will also have a general coordination mechanism through which resources will be distributed and shared among its subprojects.

The resources and language models developed for each task will be made available to the public, thus allowing their use in all kinds of applications, services, and products, by the scientific community, companies, institutions, and society in general. The results will be disseminated through a repository available at the project's web portal (which can be hosted on internal servers), as well as other established and internationally recognized repositories, such as HuggingFace, GitHub, Zenodo, etc.

Finally, the project contemplates the complete development of applications based on these

resources, which will act as visible and accessible demonstrators of the developed technology and will produce a tractor effect that will lead to the development of new products.

5. Conclusion and future work

Among the initial results of *Nós*, we can highlight the first crawl of a web-based Galician corpus and a language model based on the CCNet tools and data (Ortega et al., 2022a), and the development and testing of a Spanish-Galician neural machine translation (NMT) system prototype (Ortega et al., 2022b).

For the current year, *Proxecto Nós* aims to keep generating linguistic and computational resources to explore different subprojects. Specifically, in the first half of 2022 work will be carried out on the design of a high-quality speech corpus of sufficient size so as to allow training TTS state-of-the-art models, to be released in the last trimester. The second half of the year will also see the publication of a speech corpus for ASR. In the same timeframe, the project will publish several text corpora: parallel Galician-Spanish, Galician-English, and Galician-Portuguese corpora; a web-scale Galician text corpus, larger than the one already compiled, to be used in all the subprojects working with written text included in *Nós*; and a domain-specific corpus for automatic text generation. Based on these resources, new language models will be developed using different state-of-the-art techniques, as well as demonstrators or prototypes of a TTS system, NMT system, and automatic text generator for Galician. At the same time, throughout 2022 efforts will focus on extending and improving the first systems developed, and on validating the results obtained via the creation of high-quality gold standards.

Acknowledgements

This research was funded by the project "*Nós*: Galician in the society and economy of artificial intelligence" (*Proxecto Nós: O galego na sociedade e economía da intelixencia artificial* 2021-CP080), agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

References

[1] I. Aldabe, G. Rehm, G. Rigau, A. Way, Report on existing strategic documents and projects in LT/AI, European Language Equality (ELE), 2021.

[2] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common Voice: A Massively-Multilingual Speech Corpus, in: Proceedings of LREC 2020.

[3] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv, 2020, pp. 1–19. doi: 10.48550/arXiv.2006.11477

[4] I. de-Dios-Flores, C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. García, P. Gamallo, E. Fernández Rei, A. Bugarín-Diz, M. González González, S. Barro, X. L. Regueira, The Nós Project: Opening routes for the Galician language in the field of language technologies, in: Proceedings of the TDLE Workshop @LREC2022, pp. 52–61 Marseille, 20 June 2022.

[5] C. García Mateo, M. Arza Rodríguez (auth.), G. Rehm, H. Uszkoreit (eds.), The Galician Language in the Digital Age, Springer-Verlag, Berlin Heidelberg, 2012.

[6] B. Külebi, A. Öktem, Building an Open Source Automatic Speech Recognition System for Catalan, in: IberSPEECH, Barcelona, Spain, 2018, pp. 25–29.

[7] B. Külebi, A. Öktem, A. Peiró-Lilja, S. Pascual, M. Farrús, CATOTRON - A Neural Text-To-Speech System in Catalan. In: Proceedings of Interspeech 2020.

[8] O. Kjartansson, A. Gutkin, A. Butryna, I. Demirsahin, C. Rivera, Open-Source High Quality Speech Datasets for Basque, Catalan and Galician, in: Proceedings of the 1st Joint Workshop on SLTU and CCURL, Marseille, France, 2020, pp. 21–27.

[9] J. E. Ortega, I. de Dios Flores, P. Gamallo, J. R. Pichel, A Neural Machine Translation System for Spanish to Galician through Portuguese Transliteration, in: PROPOR 2022, Fortaleza, Brazil.

[10] J. E. Ortega, I. de Dios Flores, J. R. Pichel, P. Gamallo, Revisiting CCNet for Quality Measurements in Galician, in: PROPOR 2022, Fortaleza, Brazil.

[11] J. M. Ramírez Sánchez, C. García Mateo (auth.), M. Giagkou, S. Piperidis, G. Rehm,

J. Dunne (eds.), Report on the Galician Language (Deliverable D1.15), ELE, 2022.
[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, , Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, Y. Wu, Natural TTS Synthesis By Conditioning Wavenet On Mel Spectrogram Predictions, in: Proceedings of ICASSP, 2018.
[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, et al., Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in NLP: System Demonstrations, 2020, pp. 38–45.