Sanivert: Transformers-based End2End System for Speech Recognition in Spanish, Catalan and Portuguese in Healthcare

Pedro José Vivancos-Vicente[1], Juan Salvador Castejón-Garrido[1], Ronghao Pan[2], Camilo Caparrós-Laiz[2], José Antonio García-Díaz[2] and Rafael Valencia-García[2]

[1]VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain
[2]Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Abstract

This project focuses on improving VÓCALI's Automatic Speech Recognition systems, especially in the healthcare domain. Despite the impressive performance of state-of-the-art models such as Whisper in speech recognition, the lack of healthcare-specific training data hinders their effectiveness in this crucial domain. In addition, the problem of unpunctuated output in most of the ASR systems reduces readability, especially in scenarios with ambiguous interpretation. To address these challenges, the Sanivert project aims to adapt speech recognition models using end-2-end deep learning approaches and to develop post-processing systems for punctuation and capitalization restoration, which are crucial for improving the quality of speech recognition output. In addition, the project incorporates information extraction techniques such as named entity recognition and relation extraction to facilitate the extraction of clinical knowledge from dictated reports. With a focus on Spanish, Catalan and Portuguese, the project aligns with VÓCALI's existing solutions while strategically expanding into new markets. Ultimately, VÓCALI aims to create more adaptable and accurate ASR systems tailored to different languages and clinical specialties, ensuring improved performance in healthcare and other domains.

Keywords

ASR, punctuation restoration, knowledge extraction, natural language processing, medical domain

1. Introduction and main objective

This project is funded by the Spanish Government and the Digital Transformation Ministry and by the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia", under the 2021 call of research projects in Artificial Intelligence and other digital technologies and their integration in value chains.

Currently, VÓCALI specializes in the development of Natural Language Processing (NLP) and Automatic Speech Recognition (ASR) systems in various domains such as home automation, robotics, telephony, or public administration, with healthcare being the company's largest market niche.

With the rapid development of Transformers and pre-trained models, ASR models such as Conformer [1], Wav2Vec 2.0 [2], HuBERT [3] or Whisper [4] among others have demonstrated very good performance due to their transfer learning capabilities, which use prior knowledge learned during the pre-training phase and transfer this knowledge to specific tasks with relatively little additional training data.

Among the aforementioned models, Whisper [4] stands out because it has demonstrated human-level performance. It is a model pretrained in ASR on 680k hours of weakly supervised data and is able to generalize with 50% more robustness and fewer errors in zero-shot performance, i.e. the ability to perform speech transcription tasks without having been specifically trained on those tasks or without needing explicit training examples for each specific task. However, due to data protection and patient privacy issues, the training set does not include audio from the healthcare sector and the results for this sector is not good enough for commercial purposes. Another limitation of most ASR systems is that they produce unpunctuated output, which significantly reduces readability and overall comprehension, especially in scenarios where interpretation is ambiguous. Therefore, the restoration of punctuation and capitalization is one of the most important post-processing tasks in ASR systems [5].

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); juans.castejon@vocali.net (J. S. Castejón-Garrido); ronghao.pan@um.es (R. Pan); camilo.caparrosl@um.es (C. Caparrós-Laiz); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)
https://www.vocali.net/ (P. J. Vivancos-Vicente); https://www.vocali.net/ (J. S. Castejón-Garrido); https://github.com/Smolky (J. A. García-Díaz); https://webs.um.es/valencia (R. Valencia-García)
0009-0008-7317-7145 (R. Pan); 0000-0002-5191-7500 (C. Caparrós-Laiz); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

On the other hand, in the medical field, physicians face the challenge of managing large volumes of narrative documents containing critical patient information. Information extraction techniques such as named entity recognition and relation extraction play a key role in structuring and extracting valuable insights from clinical reports[6].

The main objective of Sanivert is to adapt and improve the VÓCALI's ASR models to incorporate Deep Learning technologies based on end-to-end (E2E) technologies, so that the systems and models generated do not depend on the isolated development of acoustic, phonological and linguistic models as is the case with current systems. This greatly facilitates the adaptation of current solutions to new languages and new clinical specialties. In addition, scoring and capitalization retrieval systems have been developed, as well as systems for extracting clinical knowledge from dictated reports.

This main objective is divided into 5 objectives.

- (OB1) Generation of linguistic resources by medical specialty and language.
- (OB2) Generation of E2E models for ASR based on Transformers.
- (OB3) Development of a score retrieval system.
- (OB4) Development of a clinical knowledge extraction system.
- (OB5) Integration of previous modules.

In this project, the technologies developed will be applied in three languages: Spanish, Catalan and Portuguese. The selection of Spanish is due to the fact that it is the language on which most of VÓCALI's solutions are based and therefore the best language to validate, and the selection of Catalan and Portuguese is a strategic business decision to reach Catalonia, Portugal and Brazil.

2. System architecture

The architecture of the proposed system is shown in figure 1. A brief description of each of these components is given below.

2.1. E2E ASR

As mentioned above, current ASR systems perform very well in Spanish. In fact, Whisper [4] has a very low word error rate (WER) in Spanish, reaching 4.2 in its large-v2 version for the Multilingual LibriSpeech (MLS) dataset [7]. However, the WER increases significantly in domains such as medicine, which contain very specialized terms that the system does not understand correctly. Table 1 shows an example of some of these words in Spanish, Catalan and Portuguese.

To address these issues, fine-tuning was performed for each language using a proprietary dataset of audio pairs and their transcriptions to adapt the Whisper versions to the medical domain. It should be noted that other existing datasets related of clinical report were used to expand the training set, such as CodiEsp [8], E3C [9], and MTSamples [1]. However, these datasets are not multimodal because they lack audio. For this reason, we used Text-to-Speech models such as Coqui-TTS [2] to transform the texts into audio with different real human voices. Coqui TSS includes tools for training new models and adapting existing models to any language. For Catalan, we used a model trained from scratch with three datasets: Festcat, OpenSLR69 and Common Voice v12. This model, based on Coqui TSS, is called *projecte-aina/tts-ca-coqui-vits-multispeaker*[3].

Currently, Whisper has 5 different configurations of variable size: tiny, basic, small, medium and large. An evaluation of these models is being conducted by fine-tuning them to determine which are better suited for the medical domain. In addition, other Transformers-based ASR systems have been tested with good results to move these models into production.

2.2. Punctuation and capitalization restoration

Punctuation restoration is a post-processing task in Natural Language Generation (NLG) that consists in adding capitalization and punctuation symbols to a text, as much of ASR does not provide this data, hindering language understanding and limiting the performance of automatic text classification models.

Nowadays, few models incorporate punctuation and capitalization restoration. Whisper [4], for example, incorporates both systems. To do this, Whisper processes the audio in the encoder, generating hidden states that the decoder will use to generate the tokens, restricting the output to the input text with the addition of punctuation.

Our punctuation and capitalization restoration model is based on a Transformers model used for sequence labeling for Spanish, Catalan and Portuguese [5, 10]. In addition, they are also able to restore capitalization improving the detection of medical entities. Both punctuation and capitalization recovery work with the same model.

For the training of punctuation and capitalization restoration model, we rely on the OpusParaCrawl [11], which contains data in Spanish, Catalan and Portuguese. This dataset contains a total of 17.2 million phrases and

[1] https://mtsamples.com/

[2] https://github.com/coqui-ai/TTS

[3] https://huggingface.co/projecte-aina/tts-ca-coqui-vits-multispeaker

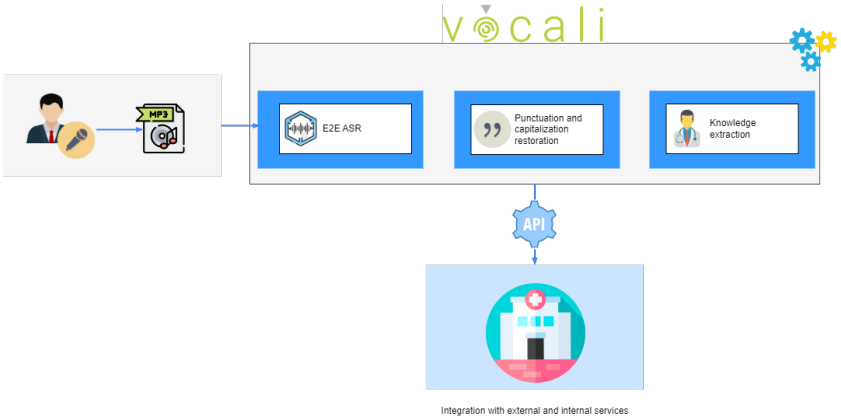

Figure 1: System architecture

Table 1
Whisper transcription error examples

Spanish	Catalan	Portuguese
recidiva	tomografia	ultrassonografia
tac	tac	ressecção
anatomopatológico	simptomatologia	antibioticoterapia
gammagrafía	gammagrafia	prednisona
parestesias	teixit	omeprazol
íleon	asimptomàtica	ipsilateral
exéresis	dexametasona	corticosteróides

774.58 million words written in Spanish. For Portuguese, we use the "pt-en" partition, which consists of 84.9 million sentences and 2.74G words written in Portuguese. Two extensions have been made to this dataset. First, we include an augmentation for Catalan based on the work described at [12], in which the authors consider replacing of words with *unknown*, random insertion and random elimination. We use this method but with a back-translation technique that consists of first translating the text into a specific language and then back-translating it into its original language. When performing this translation across different languages, translation models often replace some words with synonyms or generate new phrases with a similar meaning. Secondly, since this corpus does not contain texts related to the health sector, we have expanded it with texts from electronic clinical reports that VÓCALI has, with the aim of adapting and improving the results in the health sector.

To train these models, we first cleaned the dataset and divided it into a set of *tokens*. We then made a custom division of the dataset into training, evaluation and test sets, and, finally trained the Transformers models using the sequence tagging approach.

2.3. Knowledge extraction

The knowledge extraction system is capable of extracting and annotating natural language text with medical concepts based on the use of ontologies and entity recognition and semantic annotation technologies. This module is based on previous work of the research group[13, 14, 15]. First, medical terms are recognized using ontologies, for which we have compiled several lists of medical concepts and related them using an ontology that provides the structure from which new knowledge can be inferred. In addition, medical entities and other data are recognized in the reports. Finally, temporal expressions and quantities are detected using regular expressions.

The extracted information can be exported in standard formats to facilitate interoperability with external HIS/HCE-based systems. The HL7 FHIR [16] standard is used for this purpose. The system allows the generation of the electronic prescription using the recognized entities such as drugs, diseases, time expressions and persons. Once the data has been recognized and structured, it must be exported into the format specified by HL7 FHIR using partially instantiated templates. In addition, diagnostic tests can be generated from the data dictated

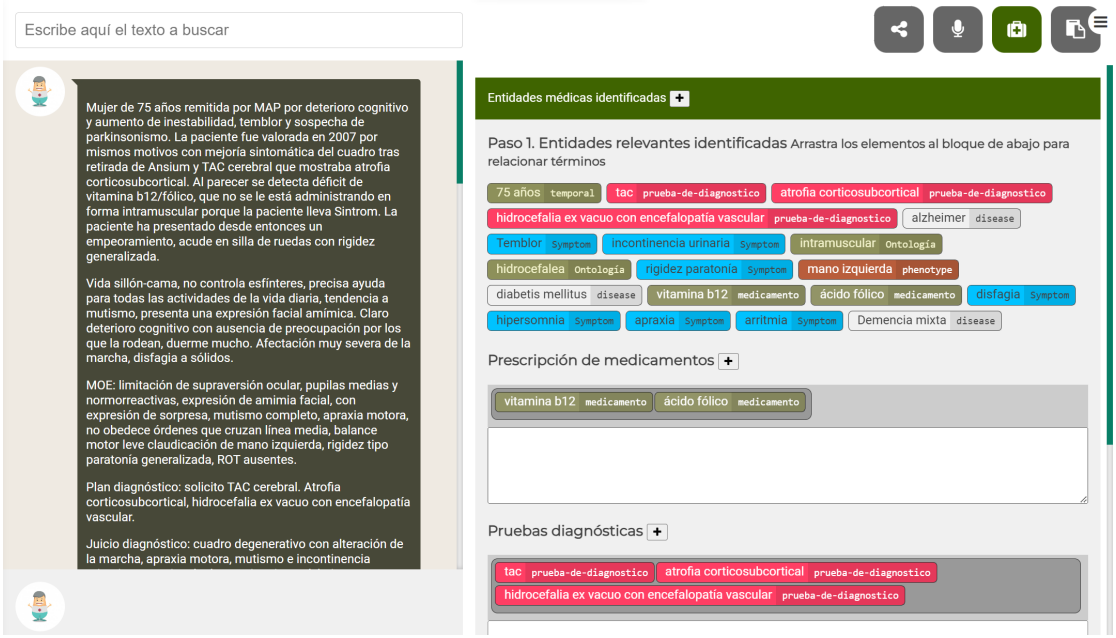

Figure 2: Web application

by the professional by adding SNOMED CT [17] codes so that the concept being treated can be easily identified.

2.4. Integration with external services

These newly developed modules are currently being integrated into VÓCALI's existing systems. This will allow us to quantify the real improvement of this process in terms of quality and process performance. Furthermore, taking into account that the models and resources developed during this work are more computationally intensive, two levels of integration systems have been implemented, one with lightweight components to provide real-time response and feedback, and another more accurate system with the rest of the functionality. In addition, to achieve the desired performance, deployment techniques based on Docker containers and dynamic per-server deployment were applied, which can dynamically scale and respond to different levels of demand.

Besides, a web application can be used to access the system and help physicians process all the information derived from the medical report. The figure 2 shows a screenshot of this interface. On the left, we can see the transcription of the medical report. On the right, we can see the identified entities, which are grouped below into sections such as diagnostic test or drug prescription.

3. Future work

The completion of this project marks a significant advancement in the development of VOCALI's ASR systems, particularly in the healthcare domain. However, there are some enhancements that will be developed in the near future.

An important direction for future work is the refinement and optimization of ASR models, with a particular focus on healthcare contexts. The acquisition and integration of more diverse and specialized healthcare audio data into the training sets would be crucial. This step aims to overcome current limitations due to privacy concerns surrounding healthcare data, thus enabling ASR models to better capture the nuances and subtleties of medical terminology and scenarios.

Improving post-processing systems for punctuation and capitalization restoration is critical, with a particular focus on overcoming the challenges posed by medical acronyms and abbreviations. Advanced algorithms tailored to accurately restore punctuation and capitalization while effectively deciphering medical terms are needed, possibly using context-aware models and domain-specific dictionaries or ontologies. These improvements are aimed at improving the readability and comprehension of transcribed clinical reports, benefiting healthcare professionals with clearer and more understandable transcriptions, and ultimately contributing to

the usability and reliability of ASR systems in medical contexts.

Finally, future work could explore the feasibility and effectiveness of adapting the ASR systems to additional languages, thereby extending the reach and applicability of the technology developed.

Acknowledgments

This work was funded by the Spanish Government, Ministerio para la Transformación Digital y la Función Pública through the "Recovery, Transformation and Resilience Plan" and also funded by the European Union NextGenerationEU/PRTR through the research project 2021/C005/0015007

References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented transformer for speech recognition, 2020. arXiv:2005.08100.

[2] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. arXiv:2006.11477.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. arXiv:2106.07447.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, 2022. arXiv:2212.04356.

[5] R. Pan, J. A. García-Díaz, P. J. Vivancos-Vicente, R. Valencia-García, Evaluation of transformer-based models for punctuation and capitalization restoration in catalan and galician, Procesamiento del Lenguaje Natural 70 (2023) 27–38. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6476.

[6] M. Y. Landolsi, L. Hlaoua, L. Ben Romdhane, Information extraction from electronic medical documents: state of the art and future research directions, Knowledge and Information Systems 65 (2023) 463–516. URL: https://doi.org/10.1007/s10115-022-01779-1. doi:10.1007/s10115-022-01779-1.

[7] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, R. Collobert, Mls: A large-scale multilingual dataset for speech research, in: Interspeech 2020, ISCA, 2020. URL: http://dx.doi.org/10.21437/Interspeech.2020-2826. doi:10.21437/interspeech.2020-2826.

[8] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2020.

[9] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The e3c project: Collection and annotation of a multilingual corpus of clinical cases, Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 (2020). URL: https://api.semanticscholar.org/CorpusID:229293442.

[10] R. Pan, J. A. García-Díaz, R. Valencia-García, Evaluation of transformer-based models for punctuation and capitalization restoration in spanish and portuguese, in: E. Métais, F. Meziane, V. Sugumaran, W. Manning, S. Reiff-Marganiec (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2023, pp. 243–256.

[11] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, J. Zaragoza, ParaCrawl: Web-scale acquisition of parallel corpora, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4555–4567. URL: https://aclanthology.org/2020.acl-main.417. doi:10.18653/v1/2020.acl-main.417.

[12] T. Alam, A. Khan, F. Alam, Punctuation restoration using transformer models for high-and low-resource languages, in: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), Association for Computational Linguistics, Online, 2020, pp. 132–142. URL: https://aclanthology.org/2020.wnut-1.18. doi:10.18653/v1/2020.wnut-1.18.

[13] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america, Future Generation Computer Systems 112 (2020) 641–657. URL: https://www.sciencedirect.com/science/article/pii/S0167739X2030892X. doi:https://doi.org/10.1016/j.future.2020.06.019.

[14] M. Ángel Rodríguez-García, R. Valencia-García, F. García-Sánchez, J. J. Samper-Zapater, Ontology-

based annotation and retrieval of services in the cloud, Knowledge-Based Systems 56 (2014) 15–25. URL: https://www.sciencedirect.com/science/article/pii/S0950705113003171. doi:https://doi.org/10.1016/j.knosys.2013.10.006.

[15] J. M. Ruiz-Sánchez, R. Valencia-García, J. T. Fernández-Breis, R. Martıínez-Béjar, P. Compton, An approach for incremental knowledge acquisition from text, Expert Systems with Applications 25 (2003) 77–86. URL: https://www.sciencedirect.com/science/article/pii/S0957417403000083. doi:https://doi.org/10.1016/S0957-4174(03)00008-3.

[16] R. Saripalle, C. Runyan, M. Russell, Using hl7 fhir to achieve interoperability in patient health record, Journal of biomedical informatics 94 (2019) 103188.

[17] D. Lee, R. Cornet, F. Lau, N. De Keizer, A survey of snomed ct implementations, Journal of biomedical informatics 46 (2013) 87–96.