Developing and Validating an Automatic Support System for Tumor Coding in Pathology Reports in Spanish

Fabián Villena, DDS, MSc[1,2,3]; Pablo Báez, PhD[4,5]; Sergio Peñafiel, MEng[6]; Matías Rojas, MEng[3]; Inti Paredes, MD, PhD[6]; and Jocelyn Dunstan, PhD[2,3,7]

DOI https://doi.org/10.1200/CCI.24.00124

ABSTRACT

PURPOSE Pathology reports provide valuable information for cancer registries to understand, plan, and implement strategies to mitigate the impact of cancer. However, coding essential information from unstructured reports is performed by experts in a time-consuming manual process. We developed and validated a novel two-step automatic coding system that first recognizes tumor morphology and topography mentions from free text and then suggests codes from the International Classification of Diseases for Oncology (ICD-O) in Spanish.

MATERIALS AND METHODS We created an annotated corpus of tumor morphology and topography mentions consisting of 1,101 documents. We combined it with the CANTEMIST corpus (Cancer Text Mining Shared Task). Specifically, we implemented a named entity recognition (NER) model using the bidirectional long short-term memory network-conditional random field architecture enhanced with a stacked embedding layer. We applied transfer learning from state-of-the-art pretrained language models to obtain high-quality contextual representations, thus improving the detection of entities. The mentions found using this model were subsequently coded using a search engine tailored to the ICD-O codes.

RESULTS Our NER models achieved an F1 score of 0.86 and 0.90 for tumor morphology and topography, respectively. The overall performance of our automatic coding system achieved an accuracy at five suggestions of 0.72 and 0.65 for tumor morphology and topography, respectively.

CONCLUSION These results demonstrate the feasibility of implementing natural language processing tools in the routine of a cancer center to extract and code valuable information from pathology reports. Our recommender system allows reliable and transparent coding at the moment of consultation. This publication shares the annotated corpus in Spanish, annotation guidelines, and source code to reproduce our experiments.

ACCOMPANYING CONTENT

Data Supplement

Accepted January 8, 2025
Published February 24, 2025

JCO Clin Cancer Inform
9:e2400124

INTRODUCTION

The free text represents a significant proportion of patients' medical records and has particular challenges because of the extensive use of abbreviations, the presence of negated information, the variability of clinical language between medical specialties, and the restricted availability for privacy reasons. Natural language processing (NLP) is the field of artificial intelligence that deals with the interaction between humans and machines through language. In medicine, typical applications of NLP are text classification, detection of drug interactions, and automatic codification of diseases.[1]

In oncology, NLP applications can be grouped into three general categories: case identification, staging, and outcome determination. Most of the work is focused on staging and includes extracting information on primary sites and parameters for the TNM classification staging system; extent of the primary tumor (T), the absence or presence and extent of regional lymph node metastasis (N), and the absence or presence of distant metastasis (M).[2,3]

A widely used nomenclature is the International Classification of Diseases for Oncology (ICD-O), developed by the WHO[4] to standardize the classification of tumor topography and morphology. The work presented here uses the Spanish version of the ICD-O (third version). The ICD-O topographic code describes the site of origin of the neoplasm and follows a structure of type *Cmm.s*, where *C* is the explicit character C, *mm* represents the main site as a left-padded integer, and *s* represents the subsite as an integer.

CONTEXT

Key Objective

The study aimed to automate tumor morphology and topography coding in Spanish pathology reports using natural language processing (NLP). This work introduces a system that recognizes tumor mentions and suggests corresponding International Classification of Diseases for Oncology codes.

Knowledge Generated

The system achieved high accuracy with an F1 score of 0.86 for tumor morphology and 0.90 for topography recognition. The automatic coding tool achieved an accuracy at five suggestions of 0.72 and 0.65 for tumor morphology and topography, respectively.

Relevance *(Z. Bakouny)*

This study provides a NLP-based tool to perform tumor morphology and topography recognition from pathology reports automatically in Spanish. Automation of such simple and repetitive tasks is relatively low-hanging fruit for machine-learning-based methods and promises to both streamline clinical work and spur research efforts.*

*Relevance section written by *JCO Clinical Cancer Informatics* Associate Editor Ziad Bakouny, MD, MSc.

Human experts are used to identify the morphologic and topographic tumor information and adequately assign the ICD-O codes to the reports. Coding is fundamental for cancer registries because it has epidemiologic and research purposes and serves to plan health services and efficiently allocate resources.[5] However, manual coding of pathology reports is time-consuming, resource-intensive, and error-prone, even with skilled personnel. As summarized in Figure 1, we developed an automatic support system for tumor coding in pathology reports.

Named entity recognition (NER) is a widely studied task in NLP that seeks to identify spans of text expressing references to predefined categories such as person names, locations, and organizations.[6] Neural networks have proven to build reliable NER systems without hand-crafted features or task-specific knowledge. In particular, the most effective approaches are bidirectional encoder representations from transformers-based architectures, which have achieved the state-of-the-art results in several data sets from different domains and languages, as in the works of Cui et al,[7] Sapci et al,[8] and Das et al.[9] A recently published work showed the efficacy of NLP to detect metastasis mentions in pathology reports.[10]

Regarding the coding task, there are no works applied to the Spanish language to code both tumor morphology and topography. However, the most common framework in the available literature is single pipelines to code documents as a whole using classification models, but these approaches only function to code subsets of nomenclatures. The main techniques used are end-to-end deep neural networks to classify top-n most common codes in a particular data set.[11–13] The main reason for not using the whole set of codes on these works is the extreme imbalance in the data set: specific codes might have very few or no examples and an excessive number of classes to predict. However, some works such as those of Atutxa et al,[14] Blanco et al,[15] and Nigam[16] have addressed the problem of (extreme) multilabel classification with strategies that offer remarkable results.

In terms of related work, Mendoza-Urbano et al[17] extracted 20 cancer characteristics, including tumor topography and morphology from oncology pathology reports, using a thesaurus-based algorithm. The average fuzzy matching score was 68.3 for topography and 89.5 for morphology. Miranda-Escalada et al[18] released the first corpus of tumor morphology mentions, manually mapped to the latest Spanish version of ICD-O. This was used in the CANTEMIST track, which included three subtasks for the automatic detection of tumor morphology mentions, concept normalization, and the assignment of ICD-O codes to each mention in clinical cases. The best results were an F1 score of 0.87, an F1 score of 0.825, and a mean average precision of 0.847 for the NER, normalization, and coding, respectively. Moreover, López-Garcia et al[19] evaluated three transformer-based models with transfer learning and an ensemble strategy in the CANTEMIST-NER subtask data set, improving the previous results with an F1 score of 0.89 with their best model. Recently, Barros et al[20] obtained state-of-the-art subtasks CodiEsp-D (diseases) and CodiEsp-P (procedures) using a divide and conquer approach.

However, most coding approaches have been developed using English language pathology reports and addressing automatic ICD-O classification at the document level. Some examples are Kavuluru et al[21] who classified 57 topographic sites using support vector machines, Qiu et al,[22] who automatically extracted six topographic codes in breast and lung cancer using a convolutional neural network (CNN),

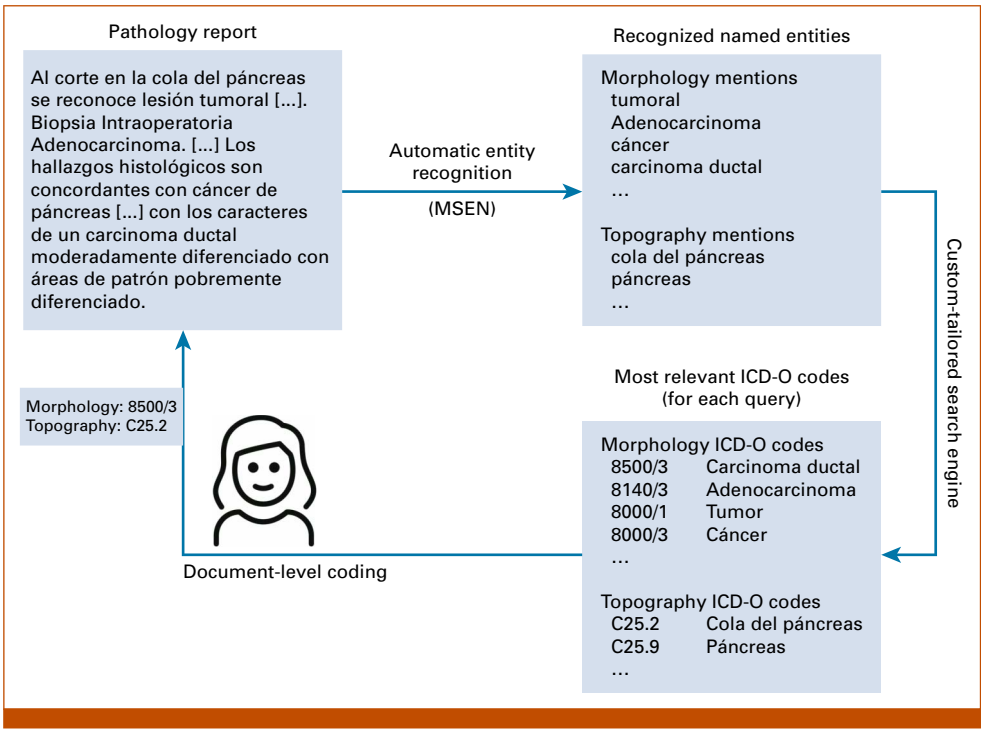

FIG 1. Schematic workflow overview for the proposed automatic support system for tumor coding in biopsy reports. It starts from the free text and continues with identifying entities and the subsequent suggestion of codes while showing the ICD-O description. Finally, the person in charge can decide on the appropriate codes. The pathology report shown in the picture can be translated into English as follows: The cut in the tail of the pancreas shows a tumor lesion [...]. Intraoperative biopsy adenocarcinoma. [...] Histological findings are consistent with pancreatic cancer [...] with the characteristics of a moderately differentiated ductal carcinoma with areas of poorly differentiated pattern. ICD-O, International Classification of Diseases for Oncology; MSEN, multiple single entity NER; NER, named entity recognition.

Wadia et al[23] who detect suspicious lung cancer from computed tomography reports, Gao et al[24] implemented models based on hierarchical attention networks to extract 12 topographic codes, Saib et al[25] present two hierarchical CNN classification methods to predict the codes of nine morphological and eight topographical classes in breast cancer case[26] and Alawad et al[26] used a multitask CNN to classify 65 primary cancer site classes. Dai et al[27] extracted nine cancer registry concepts from colorectal cancer pathology reports that were used to code eight items, including tumor morphology. Unlike the previous classification-based works, they used a hybrid strategy combining a deep sequential labeling neural network with an expert system. Recently, Wang et al[28] used a bidirectional recurrent neural network to code in ICD.

We developed and validated a novel two-step automatic coding system that suggests codes from the ICD-O in Spanish to streamline the coding processes performed by humans.

MATERIALS AND METHODS

This study was conducted at the Fundación Arturo López Pérez (FALP), the largest cancer center in Chile, with the approval of the Institutional Review Board. The center obtained informed consent from their patients. For reproducibility, we release the code we used to perform the experiments detailed in the following sections at GitHub.[29]

Data Set

We collected 1,101 anonymized pathology reports from the FALP cancer registry written in Spanish by two pathologists between the years 2017 and 2021. The entire set of reports already came with ICD-O morphologic and topographic codes assigned at the document level by FALP expert coders, without missing data. Subsequently, all tumor morphology and topography mentions inside the documents were manually annotated by clinical experts as described in the Data Supplement. We will refer to this resource as the *FALP corpus*.

In addition, we used the *CANTEMIST corpus*,[18] an open annotated corpus developed at the Barcelona Supercomputing Center in Spain. This corpus comprises 1,301 oncologic clinical case reports written in Spanish and manually annotated by clinical experts with mentions of tumor morphology linked to an ICD-O code. Table 1 shows the characteristics of the FALP and CANTEMIST corpora.

TABLE 1. Corpus Statistics for the FALP and CANTEMIST Corpora

Attribute	FALP	CANTEMIST	Total
Documents	1,101	1,301	2,402
Tokens	182,088	1,124,887[a]	1,306,975
Vocabulary	5,266	29,197[a]	32,255
M-annotations	9,671	16,030	25,701
T-annotations	6,988	—	6,988
M-unique codes	226	537	615
T-unique codes	153	—	153
Annotated tokens	27,042	37,140	64,182
Mean entities per document	15.1	12.1	13.5
Mean (STD) of tokens per entity	2.3 (2.1)	1.7 (1.2)	2.0 (1.8)

NOTE. M and T stand for tumor morphology and topography, respectively.
Abbreviations: FALP, Fundación Arturo López Pérez; STD, standard deviation.
[a]Result computed by the authors of this publication; it may differ from source publication.

Regarding preprocessing, since both corpora contain morphology mentions, we merged them to increase the amount of training data for the tumor morphology model. By contrast, the topography model was trained only with the FALP corpus. Both data sets were randomly divided into train (80%), validation (10%), and test (10%) subsets.

Automatic Entity Recognition

The detection of morphology and topography mentions was addressed as an NER task by using the multiple single entity NER[30-32] architecture. We decided on this approach since it leverages contextual and character-level embeddings to achieve outstanding performance in NER, especially when texts have misspelled and out-of-vocabulary words, which is very common in clinical texts.

Figure 2 shows a representation of the architecture. The module consists of four components: (1) an input layer that receives report tokens, (2) a stacked embedding layer to represent these tokens, (3) a bidirectional long short-term memory network encoding layer to obtain long contextual information, and, finally, (4) the conditional random field and Viterbi algorithms to decode the most probable tag sequence using the IOB2 tagging format. For additional details of each component, check the study by Rojas et al.[32] The experiments were performed on an NVIDIA Tesla V100 GPU with 32 GB of video random access memory, and the training of both models took a total of 28 hours. We

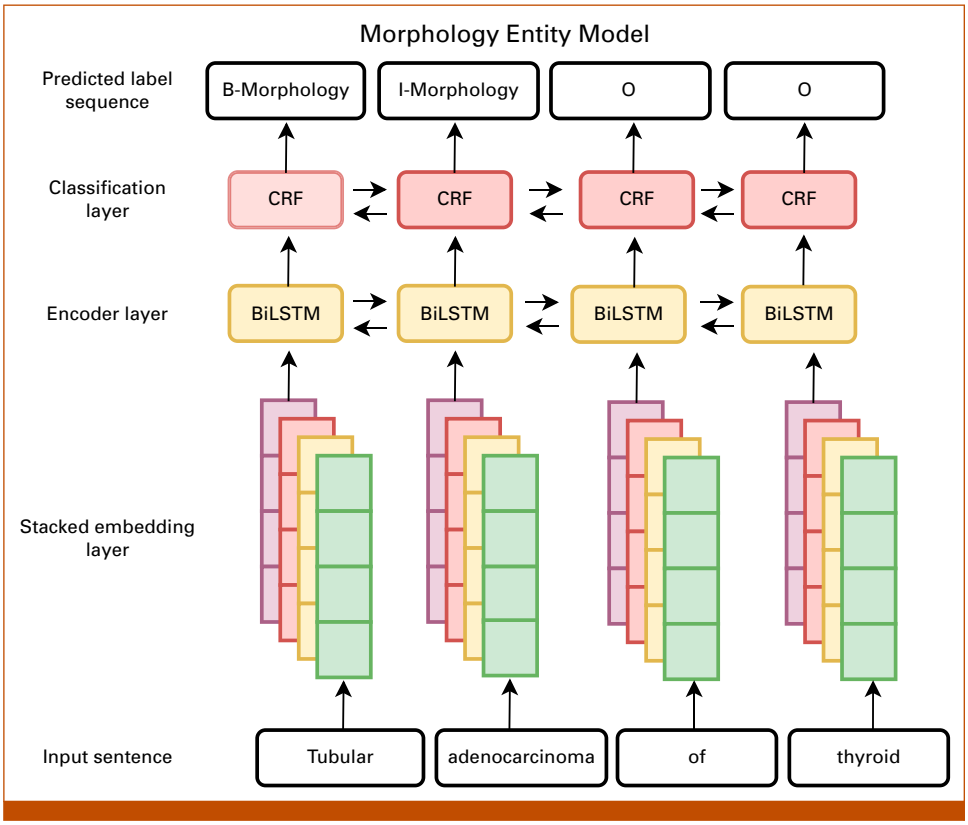

FIG 2. Overview of the NER module in the MSEN model. The figure shows, as an example, the detailed NER architecture of the morphology entity. The topography model is identical, and the results are combined in the end. BiLSTM, bidirectional long short-term memory network; CRF, conditional random field; MSEN, multiple single entity NER; NER, named entity recognition.

performed a five-fold cross-validation to validate the results and reported the average performance.

Automatic Coding

We indexed each ICD-O code using the Whoosh Python Search Engine Framework (Whoosh 2.7.4 documentation[33]) and scored each query over the index using the BM25 metric. In the study by Wijewickrema et al,[34] an exhaustive study showed the advantages of using this metric over others, such as cosine similarity. This metric ranks the index C of codes c on the basis of the query Q, consisting of words $q_1, \ldots, q_n$ using a score given by

$$\text{score}_{\text{BM25}}(c, C, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, c) \cdot (k_1 + 1)}{f(q_i, c) + k_1 \cdot \left(1 - b + b \cdot \frac{|c|}{\text{avgdl}(C)}\right)},$$

where $f(q_i, c)$ is the frequency of appearance of q_i on c of the query Q, $|c|$ is the size of the code c in words, and avgdl (C) is the average size of the codes of the index C. The hyperparameters k_1 and b are fixed on 0.75 and 1.2, respectively, on the basis of the original implementation.[35] IDF (q_i) is the Inverse Document Frequency of the word q_i defined by

$$\text{IDF}(q_i) = \log 10 \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

with N being the size of the index C and $n(q_i)$ being the number of codes that contains word q_i.

To obtain contextual information about the task we are solving, we analyzed the global distribution of codes in our data set to weight the BM25 scores by the prior global probability of appearance of each code on our data set. The prior probability score was defined by

$$\text{score}_{\text{prior}}(c, D) = \frac{f(c, D)}{|D|},$$

where $f(c, D)$ is the frequency of appearance of code c on data set D and $|D|$ is the size of the data set D. Particularly, for determining the prior probabilities on our task, we used only the distribution of codes on the training subset. The final score for the ranking of the codes given a query is given by

$$\text{score}(c, C, Q, D) = \text{score}_{\text{BM25}}(c, C, Q) \cdot (\text{score}_{\text{prior}}(c, D) + 1).$$

The function $\text{score}_{\text{prior}}$ outputs values between 0 and 1, but we added 1 to the result for convenience, so it positively modulates the output of the final score. This is important for codes with priors of 0 or very close because regardless of the BM25 result, the product would be 0 or very low when calculating the final score.

The proposed system's performance was measured using (1) the accuracy at k suggestions (Accuracy@ k), considering a suggestion correct if the true code was contained in the suggested list of codes, and (2) the mean reciprocal rank (MRR), estimated for the suggestions at multiple k (MRR@ k) as the multiplicative inverse of the rank of the true code in the suggested list of codes. Both performance metrics were calculated over the test subset. An extensive analysis of errors in the NER and coding steps is presented in the Data Supplement.

RESULTS

The FALP Corpus

We annotated a corpus comprising 1,101 documents, resulting in 9,671 morphology and 6,988 topography mentions (see Table 1 for details). This task was performed by two expert annotators who achieved an exceptional level of interannotator agreement, with an F1 score exceeding 0.9. A detailed description of the annotation protocol is provided in the Data Supplement.

An examination of the annotated corpus revealed a diverse distribution of codes, comprising 226 unique morphology codes and 153 unique topography codes. Notably, a significant imbalance characterizes the distribution of codes, where a small subset of codes dominates most of the data set. Specifically, a mere 21 codes account for approximately 80% of the entire morphology code set, whereas a similarly small subset of 23 codes represents 80% of the entire topography code set. The Data Supplement (Figs S12 and S13) also shows this imbalance.

Automatic Entity Recognition

As shown in Table 2, the entity recognition model for tumor topography achieved the highest performance, with an F1 score of 0.895, despite being trained on a smaller number of examples. By contrast, the tumor morphology model achieved an F1 score of 0.857. Both scores represent the average result obtained from five-fold cross-validation, reflecting the stability and generalization of the model on different data partitions.

Automatic Coding

The performance metrics for automatic coding were computed on the test subset for values of k from 1 to 30. For a k = 5, the accuracy and MRR of morphology increased 0.13 and 0.10 units, respectively, when using priors, whereas for

TABLE 2. Performance Metrics for the Automatic Entity Recognition Model Over Each Test Subset

Tumor Entity	Precision	Recall	F1 Score	Support
Morphology	0.838	0.876	0.857	1,933
Topography	0.899	0.891	0.895	350

NOTE. Support is the amount of true annotations on the subset. The reported results are an average over a five-fold cross-validation.

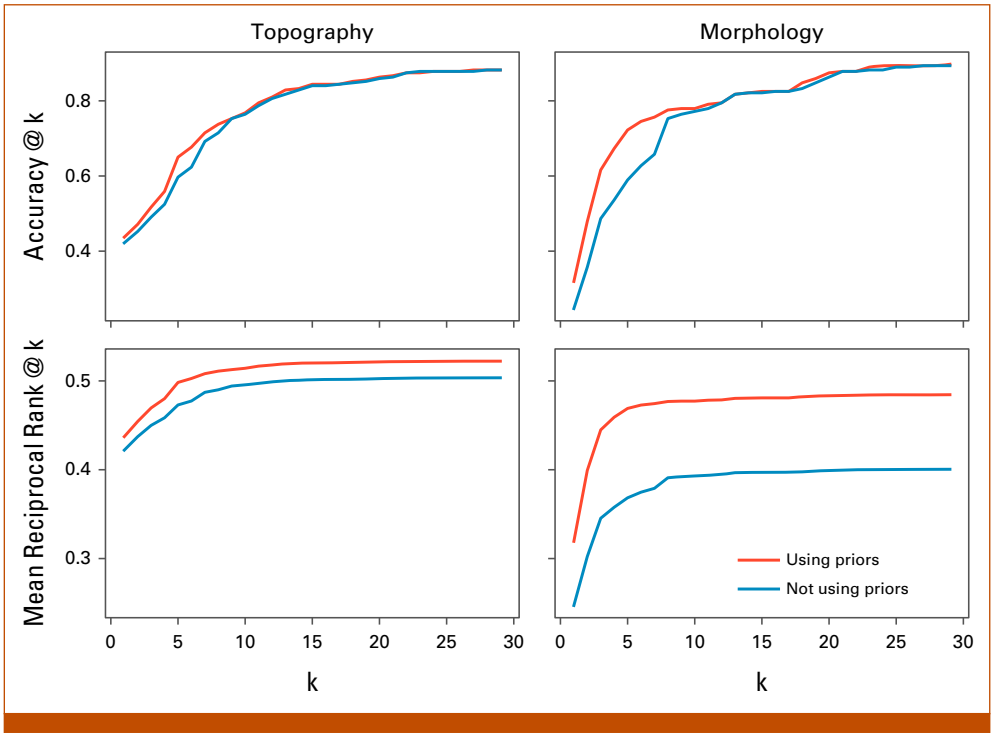

FIG 3. Performance metrics for the automatic coding system over the test subset for the complete sweep of *k*.

topography at the same *k*, the increase was 0.05 and 0.03, respectively, when using priors. The results for the complete sweep of results are reported in Figure 3, and those for specific values of *k* are reported in Table 3.

Error Analysis

Our error analysis of the automatic entity recognition model revealed 84 errors and 583 exact matches on the test subset. Most errors (53.6%) were categorized as right label, overlapping span, where the predicted mentions were classified correctly but with inaccurate boundaries. The remaining errors included false negatives (32.1%), false positives (10.7%), and wrong label, right span (3.6%). Notably, tumor topography errors were largely false negatives (48.9%), whereas tumor morphology errors were concentrated within the right label, overlapping span type (76.5%). Furthermore, our error analysis of the automatic coding model showed that it is biased toward specific responses, particularly in morphology, with overprediction of specific codes and underprediction of others. However, the posterior strategy reduced bias by 63% and 4% for morphology and topography coding, respectively. For a more detailed description of the errors, check the Data Supplement.

DISCUSSION

The detection of tumor topography mentions is easier than morphology detection, and this phenomenon translates to the performance results. This task is easier because it consists basically of detecting body parts (the site of origin of the neoplasm), which are explicitly represented inside the text spans. The model for detecting tumor topography mentions is a novelty of this work.

TABLE 3. Accuracy and MRR for the Automatic Coding System Over the Test Subset at Specific *k* Grouped by Using Priors and Not Using Priors

	Morphology				Topography			
	Accuracy		MRR		Accuracy		MRR	
k	Priors	No Priors	Priors	No Priors	Priors	No Priors	Priors	No Priors
1	0.32	0.25	0.32	0.25	0.44	0.42	0.44	0.42
5	0.72	0.59	0.47	0.37	0.65	0.60	0.50	0.47
10	0.78	0.77	0.48	0.40	0.77	0.76	0.51	0.50
20	0.87	0.86	0.48	0.40	0.86	0.86	0.52	0.50
30	0.90	0.89	0.48	0.40	0.88	0.88	0.52	0.50

Abbreviation: MRR, mean reciprocal rank.

On the other hand, performance metrics for the automatic coding were computed on the test subset for values of k from 1 to 30. For a $k = 5$, the accuracy and MRR of morphology increased 0.13 and 0.10 units, respectively, when using priors, whereas for topography at the same k, the increase was 0.05 and 0.03, respectively, when using priors. The results for the complete sweep of results are reported in Figure 3, and those for specific values of k are reported in Table 3.

Coding performance for both metrics (accuracy and MRR) is dependent on k; lower values of k explain lower performance, and higher values of k explain higher performance. The optimal value of k was determined by searching the knee of the MRR curves. Values of 6 and 4 were found for tumor morphology and topography, respectively. Therefore, we suggest the usage of $k = 5$ when implementing this recommender system. k value selection must consider the performance of the automatic coding at a specific value and the length of the suggestions list.

The overall performance of tumor morphology and topography coding increased by using the priors for the computation of the retrieval scores. The method's main advantage is the generalization of the coding step; it does not classify a subset of codes but searches over the entire index of possible codes considering local contextual information by adding the prior probabilities, improving the coding performance.

Our automatic coding model's performance in recognizing morphology mentions is comparable with that of state-of-the-art models and aligns with results reported by other authors.[18] However, our results for morphology coding are slightly lower than those reported in other studies.[18,20] We attribute this difference to the higher complexity of our task, which requires predicting a single code per document. By contrast, other tasks, such as CANTEMIST, involve predicting a set of codes, allowing for easier optimization and potentially higher performance metrics. There are no other publications to compare for the topography part of our task.

In the Data Set section, we emphasized that the morphology model was trained on the union of the FALP and CANTEMIST data sets to get more training data. However, this decision does not allow us to compare our method fairly with others previously proposed. In the Data Supplement, we present an in-depth analysis of the cross-performance of using a model trained in one corpus and applied to the other. This is particularly interesting because we are working with corpora gathered and annotated in different Spanish-speaking countries, which has not been reported previously.

This article presents a support system that facilitates the coding of tumor morphology and topography in pathology reports, an essential task for cancer registries. The system first detects and extracts mentions of morphology and topography from the text and then uses them to assign a code to each one using a tailored search on the ICD-O codes' description and historical cancer registry information. Finally, the system makes a recommendation with the most likely morphology and topography codes to assign to the pathology report.

We perform automatic clinical coding using actual clinical text and in partnership with the institution that produces it instead of using clinical trials or papers as other authors have done.[18] Another originality of this work is that we used annotated oncologic corpora from two different countries, which supports the idea that specialized clinical language is very similar across different dialects of Spanish. The FALP corpus was rigorously annotated by clinical staff with expertise in ICD-O coding, supported by medical doctors. The annotations and their mapping to ICD-O coding were consistent, and the interannotator agreement increased over time.

One of the limitations of the proposed model is that it cannot detect entities that are not written continually, that is, when they appear as two different entities for NER, such as in "Histological findings are consistent with cancer, precisely a pancreatic type," where pancreatic cancer is not written continually. This is uncommon in pathology reports but could occur in other text types. Another significant limitation is the propagation of errors. It may happen that the NER model cannot correctly recognize the span of some entity, and as a consequence, the system provides an incorrect code for such entities. This error analysis was not addressed, but such errors might have a low impact on the final coding. This is because the final coding is made with a composition of each entity's coding found. Therefore, an error in detecting an entity will only have a weight relative to the number of entities found. Finally, regarding the coding model, one limitation is that we worked with ICD-O, a vocabulary smaller than the ICD-10. Nevertheless, in a recent publication, we have demonstrated that our approach accomplished the automatic coding in ICD-10 of 25 million referrals.[36]

In future work, we plan to explore changes in the entity coding model. As mentioned above, our approach uses a search engine, but an end-to-end neural network or a retrieval augmented generation-based system could be used. Another aspect to analyze is how much the performance of the model varies when used in other types of texts, such as progress notes which are generally written by a physician at the time of appointment with the patient and consequently are less formal.

The work presented here can be easily implemented in clinical settings because each step can be deployed as an independent module inside a technological infrastructure. These modules can be used in conjunction or as isolated subsystems to solve specific tasks inside the institution.

AFFILIATIONS

[1]Department of Computer Sciences, Faculty of Physical and Mathematical Sciences, Universidad de Chile, Región Metropolitana, Chile
[2]Millenium Institute for Foundational Research on Data, Región Metropolitana, Chile
[3]Center for Mathematical Modeling—CNRS IRL 2807, Faculty of Physical and Mathematical Sciences, Universidad de Chile, Región Metropolitana, Chile
[4]Tecnología Médica, Facultad de Medicina, Universidad del Desarrollo, Región Metropolitana, Chile
[5]Center of Medical Informatics and Telemedicine, Faculty of Medicine, Universidad de Chile, Región Metropolitana, Chile
[6]Unidad de Informática Médica y Data Science, Departamento de Investigación del Cáncer, Instituto Oncológico Fundación Arturo López Pérez, Región Metropolitana, Chile
[7]Department of Computer Science and Institute for Mathematical and Computing Engineering, Pontificia Universidad Católica de Chile, Región Metropolitana, Chile

CORRESPONDING AUTHOR

Jocelyn Dunstan, PhD; e-mail: jdunstan@uc.cl.

SUPPORT

Supported by FALP and CMM (FB210005). We were also supported by ANID grants Fondecyt regular 1241825 ICN17_002 (IMFD), AFB240002 (AC3E), Postdoctoral FONDECYT 3210395, and PhD Scholarship 21220200.

AUTHOR CONTRIBUTIONS

Conception and design: Fabián Villena, Sergio Peñafiel, Inti Paredes, Jocelyn Dunstan
Financial support: Jocelyn Dunstan
Administrative support: Inti Paredes
Provision of study materials or patients: Sergio Peñafiel, Jocelyn Dunstan
Collection and assembly of data: Sergio Peñafiel, Inti Paredes, Jocelyn Dunstan
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Sergio Peñafiel
Employment: Fundación Arturo López Pérez
Travel, Accommodations, Expenses: Fundación Arturo López Pérez

Inti Paredes
Employment: Arturo Lopez Perez Foundation
Leadership: Arturo Lopez Perez Foundation

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

We thank Jocelyn Garay and Gisselle Caamano for performing annotations, Matías Espinoza for preparing the data set, and Carla Cavallera, Carolina Villalobos, Rodrigo Lagos, Álvaro Riquelme, and Marcela Aguirre for their support at FALP.

REFERENCES

1. Dalianis H: Clinical Text Mining: Secondary Use of Electronic Patient Records. Cham, Switzerland, Springer Nature, 2018
2. Yim W-W, Yetisgen M, Harris WP, et al: Natural language processing in oncology: A review. JAMA Oncol 2:797-804, 2016
3. Brierley JD, Gospodarowicz MK, Wittekind C: TNM Classification of Malignant Tumours. Hoboken, NJ, John Wiley & Sons, 2017
4. Fritz A, Shanmugaratnam K: CIE-O: Clasificación Internacional de Enfermedades Para Oncología. Washington, DC, Pan American Health Org, 2003
5. Johnson CH: The SEER Program Coding and Staging Manual 2004. Washington, DC, Cancer Statistics Branch, Surveillance Research Program, Division of Cancer, 2004
6. Chinchor N, Robinson P: Appendix E: MUC-7 named entity task definition (version 3.5). Seventh Message Understanding Conference (MUC-7): Proceedings of the 7th Conference on Message Understanding, Volume 29, 1997, pp 1-21. https://aclanthology.org/M98-1028/
7. Cui L, Wu Y, Liu J, et al: Template-based named entity recognition using BART, in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Kerrville, TX, Association for Computational Linguistics, 2021, pp 1835-1845
8. Sapci AOB, Taştan Ö, Yeniterzi R: Focusing on possible named entities in active named entity label acquisition. arXiv 10.48550/arXiv.2111.03837
9. Das SSS, Katiyar A, Passonneau RJ, et al: CONTaiNER: Few-shot named entity recognition via contrastive learning. arXiv 10.48550/ARXIV.2109.07589
10. Ahumada R, Dunstan J, Rojas M, et al: Automatic detection of distant metastasis mentions in radiology reports in Spanish. JCO Clin Cancer Inform 10.1200/CCI.23.00130
11. Shi H, Xie P, Hu Z, et al: Towards automated ICD coding using deep learning. arXiv 10.48550/arXiv.1711.04075
12. Prakash A, Zhao S, Hasan SA, et al: Condensed memory networks for clinical diagnostic inferencing. Proceedings of the AAAI Conference on Artificial Intelligence, Volume 31, 2017, pp 3274-3280
13. Cataldo-Vivar B, Allende-Cid H, Alfaro R: Automatic classification of clinical cases using deep learning techniques. 11th International Conference of Pattern Recognition Systems (ICPRS 2021), Volume 2021, 2021, pp 97-102
14. Atutxa A, de Ilarraza AD, Gojenola K, et al: Interpretable deep learning to map diagnostic texts to ICD-10 codes. Int J Med Inf 129:49-59, 2019
15. Blanco A, Pérez A, Casillas A: Extreme multi-label ICD classification: Sensitivity to hospital service and time. IEEE Access 8:183534-183545, 2020
16. Nigam P: Applying deep learning to ICD-9 multi-label classification from medical records. Technical Report, Stanford University, 2016. https://api.semanticscholar.org/CorpusID:52061700
17. Mendoza-Urbano DM, Garcia JF, Moreno JS, et al: Automated extraction of information from free text of Spanish oncology pathology reports. Colomb Médica 54:e2035300, 2023
18. Miranda-Escalada A, Farré E, Krallinger M: Named entity recognition, concept normalization and clinical coding: Overview of the CANTEMIST track for cancer text mining in Spanish, corpus, guidelines, methods and results. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020

19. López-García G, Jerez JM, Ribelles N, et al: Detection of tumor morphology mentions in clinical reports in Spanish using transformers. International Work-Conference on Artificial Neural Networks. Springer, 2021, pp 24-35
20. Barros J, Rojas M, Dunstan J, et al: Divide and conquer: An extreme multi-label classification approach for coding diseases and procedures in Spanish. Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI), 2022, pp 138-147
21. Kavuluru R, Hands I, Durbin EB, et al: Automatic extraction of ICD-O-3 primary sites from cancer pathology reports. AMIA Summits Transl Sci Proc 2013:112-116, 2013
22. Qiu JX, Yoon HJ, Fearn PA, et al: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform 22:244-251, 2018
23. Wadia R, Akgun K, Brandt C, et al: Comparison of natural language processing and manual coding for the identification of cross-sectional imaging reports suspicious for lung cancer. JCO Clin Cancer Inform 10.1200/CCI.17.00069
24. Gao S, Young MT, Qiu JX, et al: Hierarchical attention networks for information extraction from cancer pathology reports. J Am Med Inform Assoc 25:321-330, 2018
25. Saib W, Sengeh D, Dlamini G, et al: Hierarchical deep learning ensemble to automate the classification of breast cancer pathology reports by ICD-O topography. arXiv 10.48550/arXiv.2008.12571
26. Alawad M, Gao S, Qiu JX, et al: Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc 27:89-98, 2020
27. Dai HJ, Yang YH, Wang TH, et al: Cancer registry coding via hybrid neural symbolic systems in the cross-hospital setting. IEEE Access 9:112081-112096, 2021
28. Wang C, Yao C, Chen P, et al: Artificial intelligence algorithm with ICD coding technology guided by embedded electronic medical record system in medical record information management. Microprocess Microsyst 2021:104962, 2023
29. Rojas M: plncmm/falp-coding. 2023. https://github.com/plncmm/falp-coding
30. Báez P, Bravo-Marquez F, Dunstan J, et al: Automatic extraction of nested entities in clinical referrals in Spanish. ACM Trans Comput Healthc 3:1-22, 2022
31. Báez P, Villena F, Rojas M, et al: The Chilean Waiting List corpus: A new resource for clinical named entity recognition in Spanish. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Association for Computational Linguistics, 2020, pp 291-300. https://aclanthology.org/2020.clinicalnlp-1.32/
32. Rojas M, Bravo-Marquez F, Dunstan J: Simple yet powerful: An overlooked architecture for nested named entity recognition. Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, 2022, pp 2108-2117. https://aclanthology.org/2022.coling-1.184/
33. Chaput M, Zhao M, Čihař M, et al: whoosh-community/whoosh. https://github.com/whoosh-community/whoosh
34. Wijewickrema M, Petras V, Dias N: Selecting a text similarity measure for a content-based recommender system: A comparison in two corpora. Electron Libr 37:506-527, 2019
35. Robertson SE, Jones KS: Relevance weighting of search terms. J Am Soc Inf Sci 27:129-146, 1976
36. Villena F, Rojas M, Arias F, et al: Automatic coding at scale: Design and deployment of a nationwide system for normalizing referrals in the chilean public healthcare system. Proceedings of the 5th Clinical Natural Language Processing Workshop. Association for Computational Linguistics, 2023, pp 335-343. https://aclanthology.org/2023.clinicalnlp-1.37/