Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 231-234 recibido 12-12-2022 revisado 01-02-2023 aceptado 12-02-2023

Análisis y tipificación de errores lingüísticos para una propuesta de mejora de informes médicos en español

Analysis and classification of linguistic errors for a proposal to improve medical reports in Spanish

Jésica López Hernández
TECNOMOD, Departamento de Informática y Sistemas, Facultad de Informática, Campus de Espinardo, Universidad de Murcia, 30100, Murcia (España)
jesica.lopez@um.es

Resumen: Este es un resumen de la tesis doctoral realizada por Jésica López Hernández bajo la dirección del Dr. Rafael Valencia García, la Dra. Ángela Almela Sánchez-Lafuente y el Dr. Fernando Molina Molina. La defensa tuvo lugar el día 18 de mayo de 2022 en la Facultad de Letras de la Universidad de Murcia, con un tribunal compuesto por el Dr. Pascual Cantos Gómez (Universidad de Murcia), la Dra. Gema Alcaraz Mármol (Universidad de Castilla-La Mancha) y el Dr. Mario Andrés Paredes Valverde (Instituto Tecnológico Superior de Teziutlán). La tesis obtuvo la calificación de sobresaliente *cum laude* otorgada por unanimidad y mención de doctorado internacional.
Palabras clave: detección automática de errores, análisis de errores, informes médicos, lingüística computacional, procesamiento del lenguaje natural.

Abstract: This is a summary of the Ph.D. thesis written by Jésica López Hernández at University of Murcia under the supervision of Ph.D. Rafael Valencia García, Ph.D. Ángela Almela Sánchez-Lafuente and Ph.D. Fernando Molina Molina. The author was examined on May 18th, 2022 by a committee formed by Ph.D. Pascual Cantos Gómez (University of de Murcia), Ph.D. Gema Alcaraz Mármol (University of Castilla-La Mancha) and Ph.D. Mario Andrés Paredes Valverde (Higher Technological Institute of Teziutlán). The Ph.D. thesis was awarded an excellent grade and Cum Laude honours and the international mention.
Keywords: automatic error detection, error analysis, medical reports, computational linguistics, natural language processing.

1 Introducción

Uno de los principales campos de aplicación del procesamiento del lenguaje natural es el dominio biomédico. La documentación clínica contiene información de gran valor para la investigación y la práctica sanitaria, por tanto, resulta fundamental poder emplear en este campo tecnologías basadas en procesamiento automático de datos que permitan la extracción y clasificación de información, así como la anonimización de documentos clínicos, o la interoperabilidad semántica.

En el caso de los informes médicos, debido a sus características textuales y contextuales, la presencia de errores lingüísticos es común (Lai *et al.*, 2015), lo que dificulta su tratamiento automatizado. Como consecuencia, la corrección automática se convierte en un componente de gran importancia para el procesamiento de datos en informes médicos.

Los sistemas de corrección automática a la vanguardia, como las arquitecturas basadas en redes neuronales, necesitan grandes conjuntos de datos de entrenamiento para un rendimiento óptimo. Debido a la ausencia de corpus de dominio biomédico disponibles, ha ganado relevancia la recopilación y generación artificial de errores en corpus para el entrenamiento de estos sistemas. El desarrollo de una tipología de errores a partir del estudio empírico de un corpus de informes médicos va a permitir añadir nuevos patrones para la generación de errores de forma más exhaustiva y, con ello, la creación de modelos más robustos para el procesamiento de datos en medicina.

ISSN 1135-5948 DOI 10.26342/2023-70-19

Por tanto, el **propósito principal** de la tesis doctoral es la recopilación, clasificación y análisis de errores lingüísticos presentes en informes médicos en español. Mediante un estudio exploratorio con carácter descriptivo se pretende añadir otra capa de información a los métodos de detección y corrección automática disponibles para el dominio médico.

Este objetivo principal se desglosa a su vez en una serie de **objetivos específicos** entre los que se encuentran:

- Investigar sobre el estado actual del procesamiento del lenguaje natural en el dominio médico, así como la corrección automática, tanto en el ámbito general como en el dominio específico de la medicina.
- Compilar y preprocesar el corpus de estudio a partir de la recopilación de informes médicos digitalizados de varias especialidades médicas.
- Estudiar los principales métodos de detección y corrección de errores *non-word* y *real-word*. El error *non-word* genera una palabra incorrecta en el plano ortográfico; en cambio, el error *real-word* da lugar a una palabra existente y correcta idiomáticamente, pero errónea en el contexto, por tanto, este tipo de error se manifiesta en el plano semántico o sintáctico.
- Desarrollar una herramienta de cómputo y clasificación de errores.
- Identificar, analizar y clasificar de forma sistemática los errores lingüísticos presentes en informes médicos desde un enfoque cuantitativo y cualitativo.
- Comprobar si hay diferencias significativas entre las distintas especialidades y entre los errores presentes en el dominio médico y el español general.
- Contribuir a la creación de conjuntos de datos de entrenamiento más exhaustivos, que incorporen casuísticas de errores reales de informes médicos.

2 Estructura de la tesis

La tesis consta de una primera parte teórica dedicada a la investigación sobre el estado de la cuestión; y una segunda parte práctica, de carácter fundamentalmente descriptivo, que aborda el desarrollo metodológico y el análisis de los resultados. Estas dos partes se distribuyen en los siguientes capítulos:

En el **primer capítulo** se define el marco de referencia en el que se inserta la investigación y su finalidad. Por un lado, se exponen las razones que han motivado la realización de este trabajo y, por otro, se detalla la distribución de los distintos apartados que lo componen.

En el **segundo capítulo** se abordan los fundamentos teóricos y se documenta el estado de la cuestión en lo que respecta a los dos pilares que sustentan esta investigación: la corrección automática de errores y el lenguaje médico.

En el **tercer capítulo** se explica la propuesta metodológica empleada y los experimentos desarrollados. Se define el objetivo principal de la tesis y se formulan los objetivos específicos para dar respuesta al problema de investigación planteado. En segundo lugar, se presenta el corpus objeto de estudio, se proponen los criterios de análisis que se van a tener en cuenta y las distintas convenciones en cuanto al tratamiento de los datos. En la sección dedicada al procedimiento se describen las distintas fases del enfoque metodológico llevado a cabo, que incluye el preprocesamiento del corpus, la detección y corrección de errores *non-word* y *real-word* respectivamente, y el cómputo y clasificación de los errores detectados.

El **cuarto capítulo** comprende el análisis de datos a partir de los resultados obtenidos. Se realiza un análisis cuantitativo teniendo en cuenta la frecuencia, la distancia de edición, el tipo (omisión, sustitución, inserción y transposición) y subtipo de error, y la posición del error. Por su parte, en el análisis cualitativo se realiza un desglose pormenorizado de los distintos tipos de patrones de errores detectados, mencionando otros aspectos lingüísticos que pueden ser de utilidad para la finalidad del estudio.

El **quinto capítulo** incluye las conclusiones obtenidas, así como las limitaciones y desafíos presentes en la investigación y, por último, las sugerencias de líneas de trabajo futuras que servirán para mejorar y ampliar los resultados.

Finalmente, el **último capítulo** presenta las principales aportaciones científicas derivadas de esta tesis doctoral, incluyendo los artículos de investigación, los capítulos de libro y las comunicaciones en congresos.

3 Contribuciones más importantes

A continuación, se mencionan las principales contribuciones que emanan de la tesis:

Estado de la cuestión. Se ha aportado una revisión bibliográfica (López-Hernández, Almela y Valencia-García, 2019) en torno a la corrección automática y al análisis de errores en el ámbito biosanitario, que ha incluido el estudio de las principales técnicas de detección y recursos utilizados en el área. El fin principal fue conocer los desafíos y limitaciones actuales que presentaba la corrección automática específicamente en el lenguaje médico y proporcionar una síntesis de todas las investigaciones relevantes hasta la fecha.

Corpus. El corpus objeto de estudio está constituido por una recopilación de informes médicos electrónicos en español pertenecientes a las especialidades médicas de urgencias, unidad de cuidados intensivos (UCI), psiquiatría y cirugía general. El corpus, que está anonimizado, contiene un total de 2 321 826 *tokens* y ha sido sometido a un preprocesamiento y normalización para facilitar su tratamiento y análisis. Es un corpus privado, perteneciente a la empresa Vócali (https://vocali.net/), por lo que no puede ser distribuido.

Sistema. Se ha desarrollado un sistema para la detección y corrección de errores *non-word* (ortográficos), que incluye la comparación con un lexicón, la técnica de distancia de edición mínima para la generación de candidatos y la revisión manual asistida. Posteriormente, se ha trabajado en la detección de errores *real-word* (plano semántico o sintáctico). Para ello, se ha llevado a cabo la generación de modelos lingüísticos, la representación vectorial de las palabras del corpus a partir de Word2Vec y el etiquetado gramatical del corpus.

Por último, se ha desarrollado una herramienta de cómputo y clasificación con la que se ha efectuado una categorización sistemática de los errores detectados, junto con la creación de categorías adaptadas para estos errores. Esta herramienta permite la generación de matrices de confusión, que muestran qué carácter es sustituido por otro y con qué frecuencia. Como resultado, se han identificado los errores de sustitución más comunes y las combinaciones de caracteres involucradas.

Análisis. El análisis de resultados ha permitido recopilar los tipos de errores más frecuentes, conocer si existen diferencias significativas en los resultados de las especialidades médicas analizadas o si hay diferencias entre los errores detectados en el dominio médico y la tipificación existente sobre errores del español no especializado.

- **Análisis cuantitativo**: Se han detectado un total de 76 711 errores en un corpus formado por 2 321 826 *tokens*, lo que supone una tasa de error del 3,3 %. La especialidad con un porcentaje de errores más alto es urgencias. Los resultados indican que el tipo de error que ocurre con un porcentaje mayor en todas las especialidades es el de omisión de tilde y, en segundo lugar, el de omisión de letra, y la mayor parte de los errores se producen a distancia de edición 1. La mayoría de los errores se concentran en un número limitado de pares de caracteres. Entre ellos, destacan los pares de caracteres que generan confusión por su similitud fonética y por el desconocimiento de las normas académicas que regulan su uso. Por último, se observan casos cuyo error es motivado por las posiciones adyacentes de estas letras en el teclado y que evidencian que son errores de actuación o de tipo mecánico.

  Estos resultados se reflejan en López-Hernández, Almela y Valencia-García (2021), donde se muestra una primera aproximación, con un análisis de errores *non-word* en informes de la especialidad de urgencias; y en López-Hernández y Almela (2021), que presenta los resultados tras ampliar la variabilidad del corpus e incorporar informes de UCI, cirugía general y psiquiatría, aportando un análisis cuantitativo de los tipos de errores detectados.
- **Análisis cualitativo**: Se ha realizado una catalogación y descripción cualitativa de los errores detectados, que incluye una explicación de las posibles causas de aparición. En esta sección se incluye toda aquella información lingüística complementaria que puede ser útil para el desarrollo de un módulo basado en conocimiento lingüístico y el tratamiento automatizado de los errores.

  Entre los patrones de errores *non-word* detectados destacan: el uso erróneo de tildes, la formación errónea de palabras mediante derivación y composición, la escritura errónea de extranjerismos y

nombres propios, la simplificación de grupos consonánticos, la representación gráfica de fonemas errónea, la analogía con otras formas, el uso equivocado de minúsculas y mayúsculas, la creación y uso incorrecto de abreviaturas, y el tratamiento erróneo de siglas y símbolos. En el caso de los errores *real-word*, se detectan errores de paronimia, de ausencia de concordancia gramatical, de formación errónea de palabras por fenómenos de composición y prefijación, y la presencia de formas verbales anómalas en el dominio.

En Bravo-Candel *et al.* (2021) se introducen errores sintéticos en un corpus mediante reglas, para el entrenamiento de un modelo de traducción automática neuronal *Seq2seq*. En esta publicación se utilizaron dos corpus para entrenar y probar el sistema: un corpus general con 611 millones de palabras extraídas de artículos de Wikipedia en español, y un corpus de casos clínicos recopilados a partir de tres fuentes diferentes (CodiEsp, MEDDOCAN, SPACCC) y compuesto por aproximadamente 2 millones de palabras. En López-Hernández, Molina-Molina y Almela (2022) se presentan los resultados tras haber llevado a cabo la identificación, análisis y clasificación sistemática de errores *real-word* en el corpus estudiado.

Módulo basado en conocimiento lingüístico. Incorporamos un listado con los patrones detectados y la información recopilada, que puede emplearse en distintas partes del proceso de detección y corrección automática. Esta información es utilizada para desarrollar nuevas reglas formalizables por el sistema que imiten los errores detectados y, con ello, se contribuye a la creación de conjuntos de datos sintéticos para entrenamiento. En el caso de los errores *real-word* es especialmente interesante contar con este repertorio, pues son errores que suelen pasar desapercibidos en los procesos de detección. Por tanto, al entrenar el sistema para que aprenda de la casuística de errores que hemos recopilado, este será más robusto.

Además de en la fase de generación de errores para el aumento de datos de entrenamiento, la recopilación de los fenómenos que más frecuentemente constituyen errores permite aportar información en la arquitectura de decisión y ponderación de alternativas de corrección.

Agradecimientos

Esta tesis doctoral ha sido financiada por el Ministerio de Educación, Cultura y Deporte de España a través de las Ayudas para la formación de profesorado universitario (FPU), del Programa Estatal de Promoción del Talento y su Empleabilidad, con referencia FPU16/03324. También ha sido apoyada por la Agencia Estatal de Investigación (AEI) a través del proyecto LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033).

Bibliografía

Bravo-Candel, D., J. López-Hernández, J. A. García-Díaz, F. Molina-Molina y F. García-Sánchez. 2021. Automatic correction of real-word errors in Spanish clinical texts. *Sensors*, 21(9):2893.

Lai, K. H., M. Topaz, F. R. Goss y L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55:188-195.

López-Hernández, J. y Á. Almela. 2021. Detección automática de errores lingüísticos en textos clínicos: análisis de patrones de error en varias especialidades médicas. *Panace@. Revista de medicina, lenguaje y traducción*, 22(53):96-108.

López-Hernández, J., Á. Almela y R. Valencia-García. 2019. Automatic spelling detection and correction in the medical domain: A systematic literature review. En *Technologies and Innovation. CITI 2019. Communications in Computer and Information Science* (vol. 1124, pp. 104-117). Cham: Springer.

López-Hernández, J., Á. Almela y R. Valencia-García. 2021. Linguistic errors in the biomedical domain: Towards an error typology for Spanish. *Sintagma: revista de lingüística*, 33:83-100.

López-Hernández, J., F. Molina-Molina y Á. Almela. 2022. Analysis of context-dependent errors in the medical domain in Spanish: a corpus-based study. *Sage Open*, 13(1).