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

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

Jésica López-Hernández
Departamento de Informática y Sistemas
Universidad de Murcia
jesica.lopez@um.es

Resumen: Los métodos actuales de detección y corrección automática de errores tienden a ser desarrollados teniendo en cuenta un enfoque puramente computacional. La solvencia de los mismos es una realidad, pero la corrección automática aún no es un problema resuelto, principalmente en los lenguajes de especialidad. En los textos del dominio médico, debido a su complejidad terminológica y particularidades lingüísticas, los correctores ortográficos usuales resultan ineficaces y el procesamiento automático supone un reto en muchas ocasiones. Por tanto, este proyecto surge con la intención de aportar un módulo basado en conocimiento lingüístico que pueda añadir otra capa de información a los métodos actuales y, en consecuencia, contribuir a la mejora de corpus pertenecientes al ámbito de la medicina. Con el fin de investigar sobre el objetivo mencionado, se diseña un corpus de estudio constituido por una recopilación de informes clínicos electrónicos, se utilizan herramientas de extracción de errores y análisis estadístico para diseñar una tipología de error, y se aplican técnicas de procesamiento de lenguaje natural y métodos de aprendizaje automático para la implementación de la propuesta.
Palabras clave: Corrección automática de errores, detección automática de errores, tipología de error, lenguaje médico.

Abstract: Current methods of automatic error detection and correction tend to be developed with a purely computational approach. Their efficacy is a reality, but automatic correction is not yet a solved problem, mainly in specialized languages. In medical domain texts, due to their terminological complexity and linguistic particularities, the usual spell-checkers are ineffective and automatic processing is a challenge on many occasions. Therefore, this project arises with the intention of providing a module based on linguistic knowledge that can add another layer of information to current techniques and, consequently, contribute to the improvement of corpus belonging to the field of medicine. In order to investigate the mentioned objective, a study corpus is designed consisting of a compilation of electronic clinical reports, error extraction tools and statistical analysis are used to design a typology of error, and natural language processing techniques and machine learning methods are applied for the implementation of the proposal.
Keywords: Automatic error correction, automatic error detection, error typology, medical language.

Lloret, E.; Saquete, E.; Martínez-Barco, P.; Sepúlveda-Torres, R. (eds.) *Proceedings of the Doctoral Symposium of the XXXV International Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, p. 38–43
Bilbao, Spain, September 25th 2019.

1 Justificación de la investigación propuesta

El desarrollo tecnológico de las últimas décadas ha permitido, mediante el procesamiento informático de datos lingüísticos, la realización de estudios y análisis de grandes corpus con mayor profundidad y eficiencia. Asimismo, en el campo de la medicina la progresiva digitalización de los registros clínicos ha ido generando mayor disponibilidad de diversos conjuntos de datos.

Las tareas de detección y corrección automática de errores son un requisito previo para cualquier tipo de procesamiento lingüístico. No obstante, la efectividad de los correctores ortográficos, como la mayoría de aplicaciones que se construyen para procesamiento de textos, depende en gran medida del dominio donde se van a aplicar.

En el dominio médico es especialmente importante que la información se presente de la forma más rigurosa y precisa posible para facilitar el proceso de comprensión, la extracción de información, la toma de decisiones, la predicción de sucesos o la interoperabilidad. Sin embargo, para el procesamiento de los documentos clínicos existen diversos inconvenientes que deben ser tenidos en cuenta: la información se presenta desestructurada y en muchas ocasiones contiene abreviaturas, ambigüedades y errores ortográficos (Ruch, Baud y Geissbühler, 2003).

Son diversos los proyectos y trabajos que investigan sobre el proceso de desambiguación y estudio de abreviaturas, siglas y acrónimos (Wong y Glance, 2011); por el contrario, no encontramos investigaciones que se centren en el estudio de los errores ortográficos en documentación clínica para el idioma español.

Nuestra experiencia previa trabajando con informes médicos procedentes de diversas especialidades nos ha permitido comprobar que estos poseen un elevado número de errores ortográficos, tipográficos y gramaticales. Los profesionales de la salud suelen sufrir sobrecarga de trabajo y disponen de poco tiempo para redactar estos documentos, por lo que no atienden a la forma, sino únicamente al contenido. Sin embargo, en un área como medicina es de especial importancia poder hacer uso de las tecnologías basadas en procesamiento automático de datos, de ahí el interés en profundizar en esta cuestión.

Es en esta realidad donde surge la pregunta que define en gran medida este trabajo: ¿qué papel puede tener la lingüística en la detección y corrección automática de errores en el campo de la medicina? En la actualidad no existen datos cuantitativos sobre patrones de error en textos que procedan del ámbito biosanitario, y tampoco hay una revisión sistemática sobre la naturaleza de los mismos. Por consiguiente, es necesario llevar a cabo un estudio y tipificación de errores que nos permita saber qué tipos de errores tienden a cometerse en este dominio, cuáles son sus propiedades y cómo podemos aportar una base de conocimiento lingüístico a los métodos de detección y corrección existentes para tal fin.

Como hemos mencionado anteriormente, los métodos actuales tienden a ser desarrollados exclusivamente desde un punto de vista computacional. Por este motivo, consideramos que un enfoque híbrido va a permitir definir rasgos de manera explícita y puede ayudar a la toma de decisiones en aquellos casos que plantean dificultades o conflictos en la elección de alternativas a la palabra errónea. A partir de este análisis de errores puede añadirse un nuevo criterio al motor de sugerencias del corrector automático, y así contribuir a tener una mayor precisión y cobertura en este dominio especializado.

2 Antecedentes y trabajo relacionado

Son múltiples los trabajos que podemos encontrar sobre detección y corrección automática de errores ortográficos si llevamos a cabo una revisión bibliográfica sistemática. Las primeras investigaciones se remontan a la década de los sesenta. En esa década se define el concepto de *distancia de Levenshtein*, que alude al número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra; y se establecen cuatro operaciones básicas de edición (Damerau, 1964; Levenshtein, 1966):

- Adición: se inserta un carácter.
- Omisión: se elimina un carácter.
- Sustitución: se elimina un carácter y se inserta otro distinto en su lugar.
- Transposición: se produce el intercambio de caracteres adyacentes.

Los métodos convencionales de corrección de errores ortográficos se basaban principalmente en el uso de diccionarios y en la

distancia de edición mínima entre un error ortográfico y sus candidatos de corrección. Con el paso de los años se han ido sumando a estos métodos nuevas técnicas, como las basadas en similitud fonética (Veronis, 1988); técnicas probabilísticas, como el análisis de n-gramas (Ahmed, Luca y Nürnberger. 2009); técnicas basadas en reglas y heurísticas (Naber, 2003); técnicas basadas en modelos de canales ruidosos o *noisy channel model* (Brill y Moore, 2000); o las más actuales basadas en aprendizaje automático y redes neuronales (Pande, 2017).

Sin embargo, la literatura sobre corrección automática en informes clínicos es mucho más limitada. En ella encontramos sistemas que incorporan diversas combinaciones de los métodos y técnicas anteriormente mencionadas, con mayor o menor tasa de éxito. Todos coinciden en señalar el importante número de errores que presentan estos textos y la complejidad del tratamiento de los registros clínicos, tanto por el gran número de abreviaturas que contienen, como por la compleja terminología, la falta de estandarización de las formas y la ausencia de revisión posterior (Patrick et al, 2010; Lai et al, 2015; Siklósi, Novák, y Prószéky, 2016; Fivez, Suster y Daelemans, 2017; entre otros).

A su vez, como adelantábamos en la anterior sección, son inexistentes los estudios realizados sobre patrones de error en documentación clínica en español, circunstancia que nos resulta llamativa. Una tipología de error es un sistema de clasificación jerárquicamente organizado para todo tipo de errores de un determinado idioma o dominio (Wedbjer Rambell, 1999). Existen distintas dimensiones posibles para clasificar los errores:

1) Non-word (palabra no existente) y real-word (palabra existente).
2) Tipo de error: sustitución, inserción, eliminación, transposición, palabra dividida…
3) Error tipográfico, ortográfico, de estilo o gramatical.
4) Posición del error.
5) Longitud de palabra.
6) Número de errores en una palabra mal escrita.
7) Error de competencia o de actuación.
8) Contexto del error.
9) Origen del error.

Se han llevado a cabo estudios sobre identificación y clasificación de errores en otros idiomas. Entre ellos, el mayor número está dedicado al inglés (Kukich, 1992; Yannakoudakis y Fawthrop, 1983; Pollock y Zamora, 1983; Mitton, 1985; Verberne, 2002; entre otros). En los últimos años también han sido publicados estudios sobre patrones de error en portugués (Gimenes, Roman y Carvalho, 2015), en húngaro (Siklósi, Novák, y Prószéky, 2016), en japonés (Baba y Suzuki, 2012), en danés (Paggio, 2000) o en punjabi (Lehal y Bhagat, 2007). Es destacable el número de tipologías y estudios sobre patrones de error desarrollados en el ámbito de aprendizaje de lenguas (Nagata, Takamura y Neubig. 2017).

En el caso del español, hallamos dos trabajos sobre tipologías enfocados a tareas de corrección automática: *Spelling Error Patterns in Spanish for Word Processing Applications* (Ramírez y López, 2006) y *Tipología de errores gramaticales para un corrector automático* (Díaz, 2005). El primero discute sobre generalizaciones previas de patrones de error en estudios realizados para otros idiomas y ofrece una nueva perspectiva sobre patrones de error en español. Es un trabajo que se enmarca en el desarrollo de un corrector para el español en Microsoft Corporation y es especialmente relevante porque se trata de la tipología más completa existente sobre errores en español. A su vez, el segundo trabajo se centra en el tratamiento de errores gramaticales y de motivación cognitiva. En él se defiende la relevancia de la creación de una tipología de error para diseñar un corrector gramatical y de estilo.

3 Descripción de la investigación propuesta

El objetivo principal de este proyecto es desarrollar una propuesta de mejora de corpus lingüísticos, pertenecientes al ámbito de la medicina, a partir del diseño de un módulo basado en conocimiento lingüístico que se combine con otras técnicas de corrección. Este objetivo general se desglosa en una serie de objetivos menores o específicos entre los que destacan:

- Compilación de un corpus de estudio a partir de informes médicos.
- Selección de metodología y criterios de análisis de error.

- Identificación, análisis y clasificación sistemática de errores contenidos en informes médicos.
- Diseño de una tipología de error y desarrollo de un modelo de error.
- Incorporación de la tipología o reglas basadas en conocimiento a herramientas y procesos de detección y corrección automática.
- Análisis, desarrollo y evaluación del prototipo.

Por tanto, nuestra hipótesis de partida defiende que el análisis de patrones de error en los textos médicos y el diseño de una tipología de error va a contribuir a la mejora de resultados en sistemas de detección y corrección automática para este dominio.

Al tratarse de un entorno muy específico, es de gran importancia en el desarrollo de la herramienta de corrección saber qué tipos de errores ocurren en los informes médicos frecuentemente y en qué contexto, para que estos se sistematicen de manera adecuada.

Llevaremos a cabo un análisis cuantitativo y cualitativo de patrones de error contenidos en los documentos clínicos, en los que se tendrán en cuenta aspectos y características tales como frecuencia de aparición, tipo de error, causa del error, posición del error en la palabra, longitud de la palabra o contexto en el que aparece. Debemos establecer unas convenciones para que nuestra tipología sea consistente, especialmente en el tratamiento que queremos dar a abreviaturas, siglas, acrónimos, anglicismos, neologismos, errores de puntuación, etc. Para tal fin, nos hemos valido de diccionarios normativos, manuales de estilo y glosarios especializados. Actualmente contamos con una primera tipología diseñada a partir del análisis de la primera muestra del corpus. Está centrada en la tipificación de errores ortográficos y va a ser ampliada progresivamente con subgrupos y nuevas especificaciones.

Asimismo, pretendemos desarrollar un modelo de error a partir del análisis de errores y la tipología. Un modelo de error puede ser aplicado en la técnica conocida como *noisy channel model* (Brill y Moore, 2000).

La última parte de la investigación es la que tiene el componente más aplicado. Se centrará en el uso de una arquitectura modular formada por diferentes bloques con distintas técnicas de corrección y en la incorporación del modelo de error al sistema, dando lugar a una combinación híbrida de criterios y contribuyendo a la ponderación de alternativas y elección de sugerencias.

4 Metodología y experimentos propuestos

Como ya hemos adelantado en la sección anterior, nuestro trabajo puede ser dividido en tres fases principales: una primera fase dedicada al estudio de la literatura, una segunda fase basada en el análisis (y de carácter eminentemente descriptivo), y una fase final en la que se desarrollará el módulo y se harán pruebas de evaluación del mismo integrado en un sistema de corrección. Los pasos que se van a llevar a cabo son los siguientes:

1. Estudio y revisión del estado del arte, delimitación del proyecto y estudio de diversas metodologías. Se comenzará con la investigación sobre técnicas de detección y corrección automática, normalización y extracción de información. Al mismo tiempo, profundizaremos en criterios y técnicas de análisis de patrones de error, además de en el estudio de las tipologías de error existentes en otros ámbitos de investigación e idiomas. Por último, llevaremos a cabo la búsqueda de recursos y herramientas existentes que son útiles para análisis lingüístico.

2. Constitución del corpus de estudio. En esta fase se llevará a cabo la recopilación de informes clínicos digitalizados, la compilación del corpus objeto de estudio y el preprocesamiento del mismo. Haremos uso de herramientas y frameworks para el procesamiento lingüístico del texto en distintos niveles: léxico, morfológico, sintáctico y semántico.

3. Análisis. Se realizará la identificación, análisis y clasificación de patrones de errores. Para obtener la lista de palabras candidatas a error utilizaremos el corrector ortográfico Hunspell, que contiene un diccionario general, recursos terminológicos, como Snomed-CT o CIE-10, nomenclaturas, lexicones, distintos listados de palabras (como listas de siglas estandarizadas), documentos y glosarios de dominio biomédico en español. Analizaremos los distintos tipos de errores y crearemos categorías adaptadas que servirán para el diseño

de la tipología de error. Además, llevaremos a cabo la búsqueda de correlación y coocurrencia entre errores. Tras la obtención de los resultados, crearemos una matriz de confusión que será utilizada para el desarrollo del modelo de error.

4. Diseño e implementación de propuesta de mejora basada en conocimiento lingüístico. En esta última fase se llevará a cabo el proceso de experimentación y evaluación del prototipo. Realizaremos una descripción de las técnicas utilizadas en nuestro sistema y de los distintos métodos de elección. Integraremos el módulo basado en conocimiento lingüístico en la arquitectura compuesta por otras técnicas y definiremos una fórmula de decisión y ponderación de las alternativas generadas para las palabras erróneas. Finalmente, llevaremos a cabo la prueba y validación del prototipo, con distintos experimentos y métricas que reflejen la cobertura real y el grado de precisión que podemos alcanzar mediante la combinación de técnicas y la integración del módulo diseñado.

5 Cuestiones de investigación

En correspondencia con el propósito de trabajo establecido, consideramos de interés plantear las siguientes preguntas de estudio:

- ¿Tiene sentido hacer una tipología de dominio? ¿Por qué no se han desarrollado más tipologías?
- ¿Es útil para el proceso de corrección automática de informes médicos contar con un estudio sobre los tipos de errores?
- ¿De qué manera el enfoque lingüístico de análisis de errores puede contribuir en nuestra propuesta de mejora de corpus médicos?
- ¿Cómo podemos incorporar el modelo de error a procesos de detección y corrección automática de errores?
- ¿De qué forma puede complementar la tipología y el análisis lingüístico a las técnicas basadas en redes neuronales?
- ¿La incorporación de un módulo basado en conocimiento junto con la combinación de técnicas y criterios de elección va a aumentar la precisión de los métodos de corrección actuales?

Agradecimientos

Esta investigación está financiada por el Ministerio de Educación y Formación Profesional de España a través del Programa Nacional de Ayudas para la Formación de Profesorado Universitario (FPU), y por la Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER / ERDF) a través del proyecto KBS4FIA (TIN2016-76323-R).

Bibliografía

Ahmed, F., E. W. Luca, y A. Nürnberger. 2009. Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. *Polibits*, 40:39–48.

Baba, Y. y H. Suzuki. 2012. How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, páginas 373–377, Jeju Island (Corea).

Brill, E. y R C. Moore. 2000. An improved error model for noisy channel spelling correction. En *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics - ACL*, páginas 286–293, Hong Kong (China).

Damerau, F.J. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Communications of ACM*, 7(3):171–176.

Díaz Villa, A. 2005. Tipología de errores gramaticales para un corrector automático, *Procesamiento del Lenguaje Natural*, 35:409–416.

Fivez P., S. Suster y W. Daelemans. 2017. Unsupervised Context-Sensitive Spelling Correction of Clinical Free-Text with Word and Character N-Gram Embeddings. En *Proceedings of the BioNLP workshop – Association for Computational Linguistics*, páginas 143–148, Vancouver (Canada).

Gimenes, P. A., N. T. Roman y A. M. Carvalho. 2015. Spelling Error Patterns in Brazilian Portuguese. *Computational Linguistics*, 41(1):175–183.

Kukich, K. 1992. Technique for automatically correcting words in text. *ACM Computing Survey,* 24(4):377–439.

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.

Lehal G. S. y M. Bhagat. 2007. Spelling Error Pattern Analysis of Punjabi Typed Text. *En Proceedings of the 2007 International Symposum on Machine Translation, NLP and TSS*, páginas 128–141.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

Mitton, R. 1987. Spelling Checkers, Spelling Correctors, and the Misspellings of Poor Spellers, *Information Processing & Management,* 23(5):495–505.

Nagata, R., H. Takamura, y G. Neubig. 2017. Adaptive Spelling Error Correction Models for Learner English. *Procedia Computer Science*, 112:474–483.

Naber D. 2003. A Rule-Based Style and Grammar Checker. Diploma thesis, University of Bielefeld.

Paggio, P. 2000. Spelling and grammar correction for Danish in SCARRIE. En *Proceedings of the Sixth Conference on Applied Natural Language Processing*, páginas 255–261, Seattle (Washington).

Pande, H. 2017. Effective search space reduction for spell correction using character neural embeddings. En *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 170–174, Valencia (España).

Patrick, J., M. Sabbagh, S. Jain y H. Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. En *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, páginas 1–8, Valletta (Malta).

Pollock, J. J. y A. Zamora. 1983. Collection and characterization of spelling errors in scientific and scholarly text, *Journal of American Society of Informatics and Science*, 34(1):51–58.

Ramírez, F. y E. López. 2006. Spelling Error Patterns in Spanish for Word Processing Applications. En *Proceedings of Fifth international conference on Language Resources and Evaluation, LREC*, páginas 93–98, Genoa (Italy).

Ruch, P., R. Baud y A. Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, *Artif. Intell. Med.* 29 (1):169–184.

Siklósi, B., A. Novák, y G. Prószéky. 2016. Context-aware correction of spelling errors in Hungarian medical documents*, Computer Speech & Language*, 35:219–233.

Verberne, S. 2002. Context-sensitive spell checking based on trigram probabilities. Master's thesis, University of Nijmegen.

Veronis, J. 1988. Computerized correction of phonographic errors. *Computers and the Humanities*, 22(1):43–56.

Wedbjer Rambell, O. 1999. Error typology for automatic proof-reading purposes. En A. Sagvall Hein, editor, *Reports from the SCARRIE project*, Uppsala.

Wong, W. y D. Glance. 2011. Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. *Artificial Intelligence in Medicine,* 53(3): 171–180.

Yannakoudakis, E. J. y D. Fawthrop. 1983. The rules of spelling errors, *Information processing and management*, 19(12):101–108.