Técnicas de Procesamiento del Lenguaje Natural para la resolución del ingreso de nuevos términos en un Servidor de Terminología

Carlos Otero[a], Santiago Wasserman[a], Daniel Luna [a], Fernando Balbachan[a], Diego D´ellera[a], Laura Gambarte[a], Fernan Gonzáles Bernaldo de Quirós[a]

[a] *Departamento de Informática en Salud, Hospital Italiano de Buenos Aires*

Resumen

Los Sistemas de Información de Salud (SIS) deben permitir obtener datos clínicos de manera estructurada para facilitar las tareas de investigación, gestión y desarrollo de sistemas de soporte para la decisión clínica sin interferir en la libertad del médico de expresar "lo que ve". Para esto existen las terminologías de interface soportadas por los servidores de terminología.
Nuestro objetivo es describir la resolución de la identificación de nuevas descripciones sobre conceptos ya ingresados al servidor de terminología del Hospital Italiano de Buenos Aires mediante técnicas de procesamiento del lenguaje natural (PLN).
Utilizando un servidor de terminología, los conceptos que requerían codificación manual (20%) fueron evaluados con técnicas PLN. De los 2964 términos pendientes (nuevas descripciones para conceptos conocidos), 2016 (68%) fueron sugerencias acertadas, 652 (22%) fueron adecuadas, 89 (3%) fueron inadecuadas y 207 (7%) fueron insatisfactorias.
Los resultados fueron satisfactorios y en nuestra experiencia quedó demostrado que la herramienta de PLN sirve además para unificar criterios de modelado de términos pendientes.

Palabras Claves: Terminología, Procesamiento del Lenguaje Natural, Sistemas de Información en Salud, Codificación, Registros Clínicos Electrónicos

Introducción

El ingreso de datos estructurados es un obstáculo para la usabilidad y la aceptación de los aplicativos por parte de los miembros del equipo de salud (1, 2). Sin embargo, los Sistemas de Información de Salud (SIS) deben capturar los datos clínicos de manera estructurada para que permitan la investigación, la gestión y el desarrollo de sistemas de soporte para la decisión clínica (3). Se han desarrollado para ese fin sistemas terminológicos para el registro sistemático de datos clínicos que relacionan entre sí los conceptos de un dominio particular, y proporcionan los conceptos relacionados y sus posibles definiciones y códigos (4). Los sistemas terminológicos se pueden construir con una terminología global, con terminologías de referencia, y terminologías de interface, cada una utilizada para diferentes propósitos. Todos estos tipos de sistemas terminológicos se pueden agrupar en un servidor de terminología y pasar del modelo básico compuesto de una lista de códigos y descripciones, a un complejo sistema de representación conceptual del vocabulario médico (5).

El Hospital Italiano de Buenos Aires (HIBA) ha desarrollado una terminología interfaz local (6) en un contexto de un servidor de terminología (7) con el objetivo de ayudar a la documentación clínica y la autocodificación de los datos clínicos en su contexto (8). El médico a través del sistema de información en salud puede ingresar la información (diagnósticos, procedimientos, fármacos, etc) libremente en los diferentes campos de ingreso, luego el servidor de terminología realiza una búsqueda en el tesauro y, de existir el término, lo autocodifica (con una performance actual del 80%) y lo almacena en su lugar correspondiente. Los términos incluidos corresponden a un concepto (entidad clínica real) y descripciones (diferentes formas de nombrar a estas entidades clínicas, sinónimos) (Figura 1). Para aumentar la tasa de reconocimiento, los términos mal ingresados por errores de tipeo, transposición de letras, abreviaturas, etc., son ingresados como sinónimos no visibles.

CONCEPTO	HIPERTENSION ARTERIAL
DESCRIPCIONES (VISIBLES)	HIPERTENSION ARTERIAL
	HIPERTENSION ARTERIAL ESCENCIAL
	HTA
DESCRIPCIONES (NO VISIBLES)	HIPERTESION ATERIAL
	HIPPRETENSION ATRERIAL
	HIPERT ART

Figura 1: Conceptos – Descripciones en el Servidor de Terminología

Todo término no reconocido es ingresado como pendiente para ser "modelado" por el equipo de codificadores e ingresar así al sistema.

El Servidor de Terminología reconoce los términos con 100% de coincidencia, si los términos varían en un carácter o por error de tipeo se produce trasposición o cambio de letras, y no

existe ese término (termino erróneo ingresado como sinónimo no visible), es generado como pendiente.

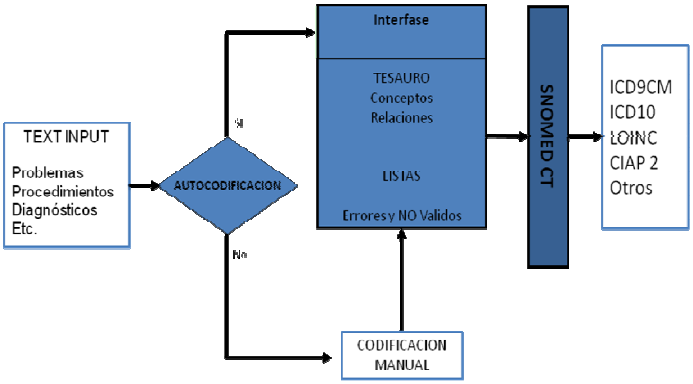

Figura 2: Estructura: Servidor de Terminología HIBA

El procesamiento del Lenguaje Natural es una disciplina innovadora, con muy poca aplicación en el ámbito de la salud en Latinoamérica, pero que ha demostrado resultados favorables en otros ámbitos (9, 10). Nos propusimos aplicar estrategias de procesamiento de lenguaje natural para reconocer similitudes entre términos y alivianar la tarea manual de codificación, mediante la sugerencia de términos "similares" al codificador.

El objetivo de este trabajo es describir la efectividad de la identificación de nuevas descripciones (nuevos sinónimos) sobre conceptos ya ingresados al Servidor de Terminología del HIBA mediante técnicas de procesamiento del lenguaje natural.

Materiales y Métodos

El Servidor Terminología del HIBA

El Servidor de Terminología de HIBA se compone de una terminología de interfaz local (tesauro) (11) que mapea a una terminología de referencia, SNOMED CT (7). El tesauro es una lista de términos creados a partir de casi 2.5 millones de entradas de texto en el repositorio de datos clínicos. Los términos incluidos en el diccionario de sinónimos se dividen en conceptos (entidades clínicas reales) y descripciones (diferentes formas de nombrar a estas entidades clínicas). Esta aplicación permite la unión de texto libre formulado por el médico en la HCE a diferentes terminologías globales, tales como la CIE-9-CM, CIE10, ICPC 2, LOINC, y otros, en forma cruzada por procesos de mapeo (Figura 2).

Esta estrategia aísla al usuario de la complejidad del uso de las clasificaciones o nomenclaturas, cuyos términos pueden no ser apropiados para el entorno, pueden tener un nivel arbitrario de detalle o puede tener demasiadas reglas para la selección de códigos. La terminología de interfaz es actualizada diariamente por un equipo de profesionales que audita el código y asocia cada nuevo término a SNOMED CT.

Cuando la ejecución del autocodificado se midió, se encontró que aproximadamente el 80% de los textos que se encuentran en una lista de problemas (6, 11) o informe de alta (12) fueron codificados automáticamente, el 20% restante requiere codificación manual.

El Servidor de Terminología esta disponible vía Web Services para otras instituciones. Estos servicios crean en primera instancia una terminología de interfaz institucional (diccionario de sinónimos) para cada organización interesada en el uso. Cada uno de estos diccionarios de sinónimos se guarda en el Servidor de Terminología del HIBA. Este proceso de replicación permite obtener la experiencia de una década de codificación en el momento de su puesta en marcha. Tras la implementación, los profesionales de codificación del HIBA auditan cada nuevo término que entra en la interfaz, y adaptan la terminología al vocabulario propio de cada institución (13, 14). Los Servicios Terminologicos que ofrece el HIBA se detallan en la Tabla 1.

Servicio	**Descripción**
Intelligent prompting	Permite una búsqueda preliminar ingresando las tres primeras letras de una palabra
Reconocimiento de Términos	búsqueda de términos ingresado y ofrecimiento de critica en línea para mejora del registro
Creación de nuevos términos	Ingreso de un Nuevo término que se enviara al sistema de auditoria
Lista de Clasificaciones	Devuelve una lista de clasificaciones
Asignación de Clasificador	Ingresado un término devuelve el código según el clasificador
Asignar DRG	Para la información ingresada, asignan un código ICD 9 y devuelve un DRG
Listar dominios	Entrega listas de Dominios disponibles
Listar Elementos de un Dominio	Devuelve términos contenidos en un dominio

Tabla 1: Servicios Terminológicos del HIBA

Resolución de la problemática

Fue necesario evaluar técnicas de Procesamiento del Lenguaje natural (PLN) que permitan otorgar al codificador "sugerencias" de términos similares a partir de los textos nuevos ingresados por el usuario.

Luego estas sugerencias fueron evaluadas como:

- **Acertadas**: si la sugerencia correspondía con el término a codificar

- **Adecuada**: si la sugerencia no era exactamente igual al término a codificar, pero por tratarse de un término de la "familia" orientaba al codificador sobre como modelar el término
- **Inadecuada**: cuando el término o los términos ofrecidos no estaban relacionados al concepto que debía ser codificado
- **Insatisfactoria**: si no traía sugerencias

Técnica utilizada

Luego de evaluar diferentes técnicas, se optó por considerar la sugerencia de términos como un problema de comparación en base a una noción de "distancia". Se define como distancia a la medida (cuantificable) en que dos términos difieren. Así, una distancia de valor 0 indica que dos términos son idénticos, mientras que valores mayores de distancia indican una creciente disimilitud. La máxima distancia está acotada por la longitud del término más extenso.
Como primer paso, se confeccionó un mecanismo para eliminar expresiones frecuentes en casos particulares pero ausentes de la terminología de referencia (p.ej., "*hace 3 años*"). Como segundo paso, se procedió a calcular diferencias. El algoritmo escogido se denomina "distancia de Levenshtein" (9), que calcula la distancia entre dos términos como la mínima cantidad de operaciones de edición (inserción, borrado y sustitución) necesarias para convertir a un término en el otro.
En la Tabla 2 se muestra que los términos se representan como secuencias de caracteres alineados, y las operaciones de edición como una matriz con los "costos" acumulados. Si bien es posible asignar un costo diferencial a cada una de estas operaciones, para penalizar ciertos cambios considerados menos frecuentes en el ámbito del que provengan los términos, pruebas preliminares para la presente implementación a partir de información en el dominio de los problemas de salud demostraron que se podían obtener iguales resultados con un valor uniforme de 1 para todas las operaciones.

		T	**U**	**N**	**O**	**R**
		1	2	3	4	5
T	1	**0**	1	2	3	4
U	2	1	**0**	1	2	3
M	3	2	1	**1**	2	3
O	4	3	2	2	**1**	2
R	5	4	3	3	2	**1**

Tabla 2: El recorrido más corto (la distancia mínima) tiene un valor acumulado de 1

Herramienta de trabajo

Los codificadores acceden al Servidor de Terminología del HIBA. La aplicación despliega los términos pendientes, o sea, aquellos que necesiten codificación. Se incorporó entonces a la izquierda de la pantalla de trabajo un nuevo campo donde se visualizan los términos sugeridos con el uso de las técnicas de PLN, y desde donde los codificadores pueden "unir" el termino en caso de que sea igual, o analizar la manera en que esta modelado si es un término de la familia. Si la sugerencia era inadecuada o insatisfactoria el codificador podía modelar por los medios habituales (Figura 3).
El codificador a su vez completaba una planilla con el resultado de las sugerencias en las categorías antes mencionadas (acertada, adecuada, inadecuada e insatisfactoria).

Figura 3: Pantalla Pendientes - Servidor de Terminología HIBA

Resultados

Entre el 3 y el 28 de enero de 2011 (20 días hábiles) 3 codificadores modelaron 2964 términos pendientes, de estos 2016 (68%) fueron sugerencias acertadas, 652 (22%) fueron adecuadas, 89 (3%) fueron inadecuadas y 207 (7%) fueron insatisfactorias.
De los 2757 términos modelados, en el 32% el sistema ofreció solo un término como sugerencia y todos correspondieron a la categoría acertada. El 49% de las veces ofreció entre 2 y 5 términos como sugerencia y el 19% más de 5 términos.
Cuando las sugerencias ofrecían más de un término, el 85% de las veces la opción correcta (ya sea por acertado o adecuado) correspondió al termino ofrecido en primer lugar.

Discusión

Si bien las técnicas de procesamiento del lenguaje natural son usadas frente a problemáticas similares en otras disciplinas, consideramos necesaria la validación de estas por tratarse de información sensible sobre la salud de los pacientes, donde un error puede tener mayores implicancias que en otros dominios.
Esta disciplina innovadora con amplio desarrollo sobre todo en países desarrollados, tiene todavía poca penetrancia en el ámbito de la salud en América latina, tal vez no solo debido a la falta de personal capacitado, sino a la poca madurez de los sistemas de información en salud. El Hospital Italiano de Buenos Aires ha incorporado a su área de terminología y documentación clínica dos lingüistas computacionales para llevar adelante proyectos en este campo.(10)
Los resultados fueron más que aceptables, pero continuamos manejando estas estrategias con auditoria "manual" por parte de los codificadores, de manera de "entrenar" y validar los

resultados, con miras a automatizar la solución y trasladarla a los usuarios.
Además de su utilidad en el reconocimiento de nuevas descripciones, de conceptos conocidos, este aplicativo sirvió de capacitador y unificador entre los codificadores, que podían analizar el modelado de términos similares y mantener la coherencia interna en el proceso de codificación.
Una limitante del trabajo es que fue solo implementado en el dominio de problemas de salud, restará validar en otros trabajos su eficacia en otros campos.
No fue un objetivo de este trabajo medir la optimización del recurso a través de la mejora de los procesos de modelado, pero estamos trabajando en un nuevo artículo que mide el "ahorro" de recursos que esta herramienta aporto al proceso de codificación y auditoria manual.

Referencias

1. Middleton B, Renner K, Leavitt M. Ambulatory practice clinical information management: problems and prospects. Healthc Inf Manage1997 Winter;11(4):97-112.

2. Rector AL. Clinical terminology: why is it so hard? Methods Inf Med1999 Dec;38(4-5):239-52.

3. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc1997 May-Jun;4(3):213-21.

4. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: Terminology and typology. Methods Inf Med2000 Mar;39(1):16-21.

5. Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp1999:42-6.

6. Osornio AL, Luna D, Gambarte ML, Gomez A, Reynoso G, de Quiros FG. Creation of a local interface terminology to SNOMED CT. Stud Health Technol Inform2007;129(Pt 1):765-9.

7. Gambarte ML, Osornio AL, Martinez M, Reynoso G, Luna D, de Quiros FG. A practical approach to advanced terminology services in health information systems. Stud Health Technol Inform2007;129(Pt 1):621-5.

8. Luna D, P. Otero, A. Gomez, M. Martinez, S. García Martí, M. Schpilberg, A. Lopez Osornio, and F.G. Bernaldo de Quiros. Implementación de una Historia Clínica Electrónica Ambulatoria: "Proyecto ITALICA". 6to Simposio de Informática en Salud - 32 JAIIO; Buenos Aires, Argentina2003.

9. Jurafsky DaJHM. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2000.

10. Were MC, Mamlin BW, Tierney WM, Wolfe B, Biondich PG. Concept dictionary creation and maintenance under resource constraints: lessons from the AMPATH Medical Record System. AMIA Annu Symp Proc2007:791-5.

11. Lopez Osornio A, S. Montenegro, S. García Martí, L. Toselli, C. Otero, I. Tavasci, D. Luna, A. Gomez, and F. Gonzalez Bernaldo de Quiros. Codificación múltiple de una lista de problemas utilizando la CIAP-2, CIE-10 y SNOMED CT. 3er Virtual Congress of Medical Informatics - Informedica2004.

12. Navas H, Osornio AL, Baum A, Gomez A, Luna D, de Quiros FG. Creation and evaluation of a terminology server for the interactive coding of discharge summaries. Stud Health Technol Inform2007;129(Pt 1):650-4.

13. Torres Casanelli C, H. Navas, S. Benitez, L. Biaggini, G. Morales, P. Navarro, D. Luna, F. Gonzalez Bernaldo de Quiros, and M. Maira. Implementación de servicios terminológicos en una red de atención ambulatoria. INFOLAC 2008 - 3er Congreso Latinoamericano de Informática Médica; Buenos Aires, Argentina2008.

14. Luna D, Lopez G, Otero C, Mauro A, Casanelli CT, de Quiros FG. Implementation of interinstitutional and transnational remote terminology services. AMIA Annu Symp Proc;2010:482-6.

Dirección para correspondencia

Dr. Carlos Martín Otero
Departamento de Informática en Salud
Hospital Italiano de Buenos Aires
carlos.otero@hospitalitaliano.org.ar