Creación de un sistema para la codificación automática de una lista de problemas

Lopez Osornio, Alejandro [a]; Luna, Daniel [a,b]; González Bernaldo de Quirós, Fernán [b]

[a] Residencia de Informática Médica, Hospital Italiano de Buenos Aires, Argentina
[b] Departamento de Información Hospitalaria, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina

Resumen

Desde 1998 en el Hospital Italiano de Buenos Aires se está codificando manualmente con la Clasificación Internacional en Atención Primaria (CIAP), en forma secundaria, los problemas médicos ingresados en su Historia Clínica Electrónica. Esto generó una base de datos con cerca de 400.000 problemas ingresados como texto libre por los médicos de cabecera y su correspondiente código CIAP seleccionado manualmente por codificadores entrenados.

Se diseñó un codificador automático que por comparación del texto que ingresa el médico con la base de datos selecciona el código CIAP más probable en base a un análisis de número repeticiones y consistencia de los codificadores. Se utilizaron técnicas de normalización del texto ingresado para su comparación.

El sistema de codificación automática codificó el 67% de los problemas médicos ingresados en un mes, con una precisión del 90%. El sistema de auto codificación además resulto ser una excelente herramienta para la auditoria de la codificación manual.

Palabras claves:

Clasificación Internacional en Atención Primaria, Codificación automática, Codificación, Lista de Problemas

Introducción

En una historia clínica electrónica (HCE) orientada a problemas la lista de problemas es la columna vertebral del modelo de información y la creación y mantenimiento de la misma es altamente recomendado por el reporte de la "Institute of Medicine (U.S.)" sobre registros médicos electrónicos [1, 2]. Un factor crítico para el éxito en la implementación de la lista de problemas es el desarrollo y mantenimiento de una codificación estandarizada de los textos narrativos ingresados por el equipo de salud en dicha lista [3, 4]. La ventaja de representar la terminología médica por medio de un sistema de codificación es que el concepto codificado puede ser fácilmente manipuleado por un sistema informático [5]. Sin embargo el texto libre no estructurado es aún hoy la forma de documentación más frecuentemente usada en medicina y por medio de la codificación del mismo se intenta disminuir la ambigüedad propia del texto narrativo. Los motivos por los cuales se codifica son múltiples tales como el económico (para facturar un acto médico), epidemiológico o estadístico (para tener datos sobre incidencia y prevalencia de patologías en una población dada), soporte para la investigación (permite la recuperación de información para estudios), asistencial (permite reclutar candidatos para programas de *disease management* y por supuesto en el contexto de una historia clínica electrónica es útil para el funcionamiento de sistemas de soporte clínico en la toma de decisiones [5].

El proceso de codificación puede ser clasificado según tres aspectos [6-8]:

Quien lo realiza: **primaria** o **secundaria**, según si el proceso de codificación lo realiza la persona responsable de brindar la atención al paciente o no. En la primaria el profesional actuante cuenta con información adicional en el momento de codificar el evento y ello le permite una mejor elección del código en relación al que elegiría un codificador "ciego" o sin contacto con el proceso de atención como es el caso de la secundaria.

Donde lo realiza: **centralizada** o **descentralizada**, según si se codifica por una unidad de codificación o en múltiples lugares simultáneamente, con lo cual el proceso de codificación recae en tantas personas diferentes como las que sean capaces de atender pacientes y es difícil lograr que frente a una misma situación el código elegido sea siempre el mismo (inconsistencias). Una respuesta a este problema es la centralización de la codificación, mediante el cual un número reducido de personas concentren el conocimiento de la clasificación a utilizar y son los responsables de asignar secundariamente los códigos correspondientes a las rúbricas de texto libre que el personal asistencial registra durante la atención.

Como se realiza: **manual** o **computarizada**, según si en el proceso de codificación interviene o no un programa informático. La codificación manual tiene desventajas tales como la persistencia de inconsistencias (variabilidad intercodificador e intracodificador), pérdida de conocimiento cuando un codificador deja de serlo, sin embargo es el modelo históricamente más utilizado.

A su vez la codificación computarizada puede ser:

- **Asistida**: en este tipo de codificación el software asiste al usuario en la asignación del código a la palabra o frase expresada en lenguaje natural, también denominados *"encoders"*. Estos sistemas van orientando al codificador por medio de preguntas u opciones que aseguran la más correcta asignación del código, con lo cual insta al usuario a buscar más detalle en el registro para la correcta asignación del código con lo cual mejora tanto la velocidad como la exactitud de la codificación [9, 10].
- **Automática**: en este tipo de codificación computarizada es el programa el que asigna el código sin intervención primaria del usuario codificador. Existen tres tipos de codificación automática [8]:
  - **Mapeo automático de términos**: los códigos son automáticamente asignados según el texto ingresado de acuerdo al mapeo que resulta de la comparación del texto ingresado con la base de conocimiento del programa. El proceso de preguntas y guías ofrecidas al usuario no son necesarios (como en la codificación asistida) ya que la lógica interna del sistema asigna los códigos automáticamente. Es importante consignar que la performance de estos sistemas son enteramente dependientes de las reglas internas que definen la lógica de asignación de códigos.
  - **Terminología controlada**: puede ser considerado como más sofisticado que el anterior. Se entiende por terminología controlada a los sistemas terminológicos diseñados para representar los datos clínicos a una granularidad consistente con la práctica diaria. Este tipo de sistemas utiliza estos vocabularios como referencia para la codificación por medio de herramientas léxicas y ontologías (UMLS, SNOMED, Read Codes).
  - **Procesamiento de lenguaje natural (PLN)**: algunos los sistemas de codificación automática pueden extraer la información clínica pertinente de los documentos y codificarlos mediante mecanismos de procesamiento del lenguaje natural.

Si bien los codificadores experimentados pueden asignar códigos comunes de memoria, también pueden asignar códigos incorrectos en forma sistemática. Expertos en registros médicos han recomendado una variedad de estrategias para mejorar la exactitud de los códigos, entre las que se encuentran el empleo de codificadores certificados, participación de los mismos en reuniones de consenso, educación permanente, creación de políticas de codificación institucionales, implementación de programas de control permanente de la calidad de la codificación y el uso de la *codificación automática* [11].

Estos sistemas de codificación automática tienen ventajas claras frente a las demás formas de codificación [6, 8, 12]:

- Incremento de la velocidad y exactitud del proceso de codificación
- Mejora la consistencia inter e intra codificador en relación a la codificación manual, por lo tanto la calidad de los códigos asignados
- El tiempo en el cual el texto libre puede estar codificado (para cualquiera de los fines para los que se codifica) se acelera drásticamente
- Elimina el cuello de botella producido en el grupo de codificadores cuando el volumen de trabajo aumenta
- Produce un ahorro en los recursos humanos destinados a la codificación
- Permite el almacenamiento del conocimiento de los codificadores entrenados en la base de datos cuando estos abandonan el grupo

La Clasificación Internacional en Atención Primaria (CIAP) [13] es un sistema de clasificación compuesto por 800 rúbricas repartidas en 17 capítulos, de muy poca granularidad para ser utilizado como terminología de referencia pero especialmente diseñado para reflejar la realidad de los diagnósticos mas frecuentes en Atención Primaria.

Desde 1998, en el sistema de Historia Clínica Electrónica Ambulatoria (HCE), los médicos de cabecera confeccionan una lista de problemas para cada paciente. La lista de problemas es una herramienta muy utilizada en la historia clínica en papel, permite que los diferentes médicos que participan de la atención de un paciente registren y organicen los diferentes problemas de salud. Los médicos ingresan los problemas mediante una línea de texto que los describen, posteriormente un equipo de codificadores asocia ese problema con un código CIAP. Aunque en la HCE siempre se muestra la descripción de texto libre el código normalizado se utiliza para el análisis estadístico y en el futuro para desencadenar recordatorios y herramientas de soporte de decisión.

En un estudio previo comprobamos la confiabilidad de esta estrategia de codificación secundaria y centralizada en forma manual en nuestro medio [14]. Reportes internacionales informan similares resultados [15].

El objetivo de este trabajo fue validar un prototipo de sistema de codificación automática por medio de mapeo automático de términos para ser incorporado a nuestra HCE. Por razones prácticas buscamos una solución fácil de implementar a un bajo costo, motivo por el cual pospusimos la intensiva creación de una fuente de ontología, la cual es un prerrequisito de todas las soluciones formales a este problema.

Materiales y Métodos

Para codificar automáticamente los problemas ingresados por los médicos decidimos crear un sistema que utilice la información de los codificadores, que durante 3 años codificaron manualmente los problemas médicos con CIAP.

El objetivo del sistema es evitar la re-codificación, es decir que si en varias oportunidades los médicos ingresaron exactamente la misma palabra y los codificadores la asociaron a un código CIAP, la próxima aparición de esta palabra se codifique automáticamente.

El sistema de codificación automática se basa en la construcción de un Tesauro que relacione los términos médicos más comunes con los códigos CIAP correspondientes. Para realizar este estudio y evaluar la utilidad de la codificación automática se realizó un procesamiento no automatizado de los datos, ejecutando cada paso de este proceso en forma manual en un sistema de bases de datos. Este tesauro se basa en una colección de términos ya codificados manualmente. Se recopilaron los datos de 3 años ingresados a través de la historia clínica ambulatoria, 390.026 registros hasta el 31/05/2002 inclusive.

Procesamiento de la lista de problemas

Esta colección se procesa en los siguientes pasos:

Normalización

El texto ingresado por los médicos para describir problemas es procesado para eliminar diferencias inherentes al formato. Esto comprende la eliminación de signos de puntuación, espacios de más, signos, conversión de todo a mayúsculas, artículos y preposiciones. Aunque algunas de estas modificaciones puedan cambiar el sentido literal de una frase, es muy poco probable que cambien el código a asignar en una clasificación de escasa granularidad como el CIAP.

Agrupación

La lista de términos es agrupada por el texto normalizado. Se cuentan la cantidad de apariciones totales del término normalizado y la cantidad de asociaciones con cada código CIAP. Ya que el trabajo de codificación humano es manual, esta sujeto a errores y no siempre el mismo texto está asociado al mismo código.

Evaluación

En base a la lista agrupada calcula la confiabilidad de la codificación humana para elegir el código a incorporar al tesauro. Por ejemplo si un término apareció 200 veces y fue asignado en 150 oportunidades al código A1 y en las 50 restantes está repartido entre los códigos A2, A3 y A4, el porcentaje de coincidencia es del 75%, el porcentaje que representa el término con mayor cantidad de apariciones.

Creación del Tesauro

Se incluyen los términos normalizados que hayan pasado al menos 10 veces por la codificación manual, para dar oportunidad a que los codificadores repitan el proceso y no asumir como correcto el primer código asignado. Y de estos se incluyen sólo los que tengan un porcentaje de coincidencia mayor al 70%, asociados al código más utilizado para describirlo. Por lo tanto el tesauro queda definido con una entrada por cada término normalizado y codificado confiablemente y el código asociado.

Remodificación automática

Una vez creado el tesauro se selecciono un período de tiempo no incluido en la confección del mismo y ya codificado manualmente, del 01/06/2002 al 10/07/2002, 8.184 registros y se utilizó como muestra para validar la codificación automática contra el modelo actual.

Se procedió luego a la codificación automática de la muestra comparando los términos normalizados con el tesauro y la asignación del código si había coincidencia. Se utilizó la coincidencia exacta de términos normalizados en lugar de estrategias avanzadas [16] para asegurar la precisión del sistema.

Resultados

El tesauro reunió a 2.171 términos ya normalizados diferentes, los cuales se habían codificado manualmente más de 10 veces y tenían una concordancia en la asignación de los códigos del 70% o más. Estos términos fueron codificados utilizando 408 códigos CIAP. Hubo 550 términos normalizados adicionales que no se incluyeron en el tesauro porque pese tener más de 10 apariciones no había un índice de concordancia suficiente en la elección del código.

Utilizando el tesauro se logró identificar y codificar automáticamente el 67% de los 5.498 problemas ingresados por los médicos durante el período de 1 mes y medio de muestra. Un 7% adicional de los términos fue reconocido por el tesauro pero no pudo ser codificado por falta de concordancia entre los codificadores manuales.

De los términos codificados, el porcentaje de coincidencia con los codificadores manuales en el mismo período fue del 89% a nivel de código CIAP completo. A nivel de aparato (la primera letra del código) la coincidencia fue del 98%.

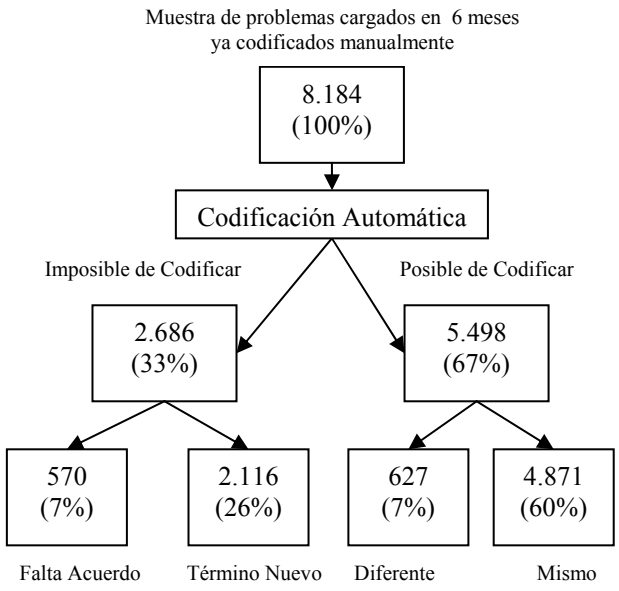

Figura 1- Resultados de la Codificación Automática

Discusión

Hay varios trabajos que reportan la utilización de sistemas de codificación automática con mapeo automático de términos [17, 18], por medio de terminología controlada [6, 19] y desarrollos mixtos [3, 4] todos utilizando múltiples sistemas de codificación y terminologías. Con respecto a los sistemas que utilizan *procesamiento de lenguaje natural* tenemos que hacer notar que hay pocos, están limitados a dominios acotados y la mayoría están desarrollados para áreas de investigación y docencia [20], la performance de algunos de estos productos ha sido testeada y obtiene similares resultados que codificadores expertos [12, 21]. Hay que tener en cuenta que este tipo de programas no están disponibles aún en idiomas diferentes al inglés donde una sintaxis más compleja dificulta el desarrollo de los mismos [9].

En un futuro trabajaremos en la creación de un servidor de terminología clínica que soporte terminologías de referencia en nuestro idioma, como la versión en español del SNOMED [22], para el procesamiento automático de la lista de problemas con la posibilidad de tener múltiples vocabularios de salida según las necesidades [23].

En el sistema de Historia Clínica Electrónica ambulatoria implementado en nuestro hospital los médicos tienen absoluta libertad al momento de ingresar el texto. Este texto es el que se presenta al abrir la historia clínica de un paciente a los médicos, y la lista de problemas es su propia herramienta de trabajo diario, por lo tanto los médicos tienden a ser muy específicos a la hora de describir un problema, con textos largos y con jergas personales. Esto aumenta la dispersión de los diferentes textos ingresados por los médicos y las posibilidades de ingreso de faltas de ortografía. Creemos que esto puede ser una posible explicación al hecho que pese a usar una base de datos de casi 400.000 registros para crear el tesauro, cargada a lo largo de 3 años y medio, en el período de la muestra el 26% de los textos ingresados son nuevos. Esta es una cantidad mayor a la reportada en estudios previos, donde por ejemplo se codificaron diagnósticos de derivaciones, donde el médico no utiliza el dato que registra [18]. Posiblemente se pueda mejorar esta dispersión con capacitación y con una modificación del sistema, donde sugiera al médico términos parecidos al ingresar un término desconocido.

Los 5.498 problemas que se codificaron automáticamente representan a lo largo de 1 mes y medio una gran cantidad de horas de codificación, con la consiguiente reducción de costos y mejoría de la calidad de la historia al tener disponible la codificación automática en la mitad de los casos.

Este sistema de codificación juega un papel muy importante también en la detección de errores humanos de codificación. En nuestra muestra se detectaron 2 tipos de errores, uno que consiste en asignar un código a un problema que no es el que le han asignado en la mayoría de las veces cuando apareció ese mismo problema con anterioridad, generalmente producidos por errores humanos o distracciones, pero puede representar la corrección de un error sistemático del equipo de codificadores. El segundo tipo de error sucede cuando un problema pasa por la codificación manual en varias oportunidades pero se le asignan diferentes códigos y ninguno de ellos llega a tener una mayoría, esto sucede con términos ambiguos o que caen en una ambigüedad de la clasificación. Sumando el 3% de problemas en el primer caso y el 9% del segundo, nos queda un 12% problemas que tienen asignado un código de dudosa utilidad.

Conclusiones

La codificación automática por mapeo automático de términos y comparación de textos normalizados es una estrategia con muchas limitaciones pero de con una gran facilidad de implementación y un costo muy reducido, lo que la hace una alternativa atractiva a la implementación de servidores de terminología en proyectos de pequeña y mediana escala. Además en sistemas de codificación manual ya en marcha, este enfoque representa una herramienta invalorable para asegurar la máxima calidad de los datos mediante la detección de errores en la codificación.

Agradecimientos

Agradecemos a todo el grupo de codificadores dependiente del Área de Informática Médica del Hospital Italiano por su constante esfuerzo y colaboración con el mantenimiento de la codificación de diferentes dominios de nuestro sistema de información.

Referencias

[1] Institute of Medicine (U.S.). Committee on Improving the Patient Record, R.S. Dick and E.B. Steen, *The computer-based patient record : an essential technology for health care*. 1991, Washington, D.C.: National Academy Press. xii, 190.

[2] Dick, R.S., E.B. Steen, D.E. Detmer and Institute of Medicine (U.S.). Committee on Improving the Patient Record, *The computer-based patient record : an essential technology for health care. Institute of Medicine (U.S.). Committee on Improving the Patient Record.* Rev. ed. 1997, Washington, D.C.: National Academy Press. xx, 234.

[3] Warren, J.J., J. Collins, C. Sorrentino and J.R. Campbell, *Just-in-time coding of the problem list in a clinical environment.* Proc AMIA Symp, 1998: p. 280-4.

[4] Campbell, J.R. and P. Elkin, *Human Interfaces: Face-to-Face with the Problem List.* Proc AMIA Symp, 1999(1-2): p. 1204.

[5] Peden, A.H., *An overview of coding and its relationship to standardized clinical terminology.* Top Health Inf Manage, 2000. **21**(2): p. 1-9.

[6] Benson, L.O., E. Kuelbs, L. Marc and C. Lock, *Implementing and evaluating computer-assisted coding of advesrse events.* Drug Inf J, 1996. **30**: p. 799-809.

[7] Beinborn, J., *Automated coding: the next step?* J Ahima, 1999. **70**(7): p. 38-43.

[8] Beinborn, J., *The automation of coding.* Top Health Inf Manage, 2000. **21**(2): p. 68-73.

[9] Hohnloser, J.H., P. Kadlec and F. Puerner, *Experiments in coding clinical information: an analysis of clinicians using a computerized coding tool.* Comput Biomed Res, 1995. **28**(5): p. 393-401.

[10] Bernstein, R.M., G.R. Hollingworth, G. Viner, J. Shearman, C. Labelle and R. Thomas, *Reliability Issues in Coding Encounters in Primary Care Using an ICPC/ICD-10-based Controlled Clinical Terminology.* Proc AMIA Symp, 1997: p. 843-7.

[11] Lloyd, S.S. and E. Layman, *The effects of automated encoders on coding accuracy and coding speed.* Top Health Inf Manage, 1997. **17**(3): p. 72-9.

[12] Morris, W.C., D.T. Heinze, H.R. Warner Jr, A. Primack, A.E. Morsch, R.E. Sheffer, M.A. Jennings, M.L. Morsch, and M.A. Jimmink, *Assessing the accuracy of an automated coding system in emergency medicine.* Proc AMIA Symp, 2000: p. 595-9.

[13] Equipo CESCA, *Clasificación Internacional en Atención Primaria (CIAP).* 1ra ed. 1990, Barcelona: Masson.

[14] Luna, D., F.G. Bernaldo de Quiros, L. Garfi, E. Soriano and M. O´Flaherty, *Reliability of secondary central coding of medical problems in primary care by non medical coders, using the International Classification of Primary Care (ICPC).* Medinfo, 2001. **10**(Pt 2): p. 300.

[15] Britt, H., *Reliability of central coding of patient reasons for encounter in general practice, using the ICPC.* Journ Informatics in Prim Care, 1998. **May**: p. 3-7.

[16] Lovis, C. and R.H. Baud, *Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language.* J Am Med Inform Assoc, 2000. **7**(4): p. 378-91.

[17] Surjan, G. and G. Heja, *Indexing of medical diagnoses by word affinity method.* Medinfo, 2001. **10**(Pt 1): p. 276-9.

[18] Letrilliart, L., C. Viboud, P.Y. Boelle and A. Flahault, *Automatic coding of reasons for hospital referral from general medicine free-text reports.* Proc AMIA Symp, 2000: p. 487-91.

[19] Franz, P., A. Zaiss, S. Schulz, U. Hahn and R. Klar, *Automated coding of diagnoses--three methods compared.* Proc AMIA Symp, 2000: p. 250-4.

[20] Lussier, Y.A., L. Shagina and C. Friedman, *Automating SNOMED coding using medical language understanding: a feasibility study.* Proc AMIA Symp, 2001: p. 418-22.

[21] Warner, H.R., Jr., *Can natural language processing aid outpatient coders?* J Ahima, 2000. **71**(8): p. 78-81; quiz 83-4.

[22] Reynoso, G.A., et al., *Development of the Spanish version of the Systematized Nomenclature of Medicine: methodology and main issues.* Proc AMIA Symp, 2000: p. 694-8.

[23] Chute, C.G., P.L. Elkin, D.D. Sherertz and M.S. Tuttle, *Desiderata for a clinical terminology server.* Proc AMIA Symp, 1999: p. 42-6.

Dirección para correspondencia

Dr. Alejandro Lopez Osornio:

alejandro.lopez@hospitalitaliano.org.ar

Residencia de Informática Médica. Área de Informática Médica. Departamento de Información Hospitalaria. Hospital Italiano de Buenos Aires. Gascón 450. Ciudad Autónoma de Buenos Aires. Argentina. (C1181ACH)