GENERACIÓN DE GRUPOS SEMÁNTICOS PARA LA CODIFICACIÓN AUTOMÁTICA DE RESPUESTAS ABIERTAS

Santana Suárez O.[1], Pérez Aguiar J.[2], Sánchez Berriel I.[3], Gutiérrez Rodríguez V.[4], Díaz Martín S.[5]
Departamento de Lenguajes y Sistemas, Universidad de Las Palmas de Gran Canaria[1,2]
Departamento de Estadística, Investigación Operativa y Computación, Universidad de La Laguna[3,4]
Escuela Técnica Superior de Ingeniería Informática, Universidad de La Laguna[5]

RESUMEN

Actualmente la explotación del contenido semántico de datos se ha convertido en un aspecto de especial relevancia en diversos campos y aplicaciones de las Tecnologías de la Información. A la hora de abordar este problema resulta obligado incorporar los conceptos implicados en la información que se procesa, lo que lleva a la definición y explotación de diccionarios, tesauros u ontologías que permitan introducir en el sistema los significados como información estratégica. En este trabajo se aborda el diseño de una herramienta capaz de extraer automáticamente conceptos implícitos en una variable de tipo texto. En todo momento el diseño ha sido abordado con un enfoque generalista que permite su uso no sólo en la codificación automática de cualquier variable de respuesta abierta, sino también en cualquier problema en que se requiera generar agrupaciones de palabras según su semántica, tales como la extracción automática de metadatos, la recuperación de información, la generación automática de resúmenes de documentos, etc. La implementación de la solución se basa principalmente en el uso combinado de tecnologías de la Lingüística Computacional y de la Minería de Textos para la construcción de grupos de términos con un nexo semántico relevante para el problema bajo estudio.

PALABRAS CLAVES

Lingüística Computacional, Recuperación de la Información, Codificación Automática, Ontologías, Minería de Textos

1. CODIFICACIÓN AUTOMÁTICA DE RESPUESTAS ABIERTAS EN ENCUESTAS

Emplear preguntas de respuesta abierta en encuestas, en las que el encuestado emplea su particular forma de expresión, aporta información valiosa a los estudios sociológicos, económicos, de mercado, de opinión, etc. Este tipo de variables genera información más enriquecedora y variada que las de respuesta cerrada, que resultan de notable valor estratégico para las organizaciones. Sin embargo, el tratamiento que se les da a las mismas requiere técnicas específicas del análisis cualitativo de datos que pueden completarse con análisis estadísticos si se codifican de forma adecuada. El uso de técnicas cuantitativas en este tipo de datos exige su codificación, lo que entraña un proceso complejo de categorización de textos: resuelto tradicionalmente de forma manual por especialistas en el área, o con la ayuda de programas apropiados. Entre las aplicaciones informáticas se encuentran las que sirven de soporte en codificación asistida (AYUDACOD)[1] o las que llevan a cabo la codificación automática (hCod, AutoCoder)[2]. Los objetivos del presente trabajo se enmarcan en este último grupo, abunda en el desarrollo de un aplicativo que permita la codificación automática de respuestas abiertas en encuestas. El proceso que se sigue se divide en los siguientes módulos (Figura 1)

- Recolección y normalización de datos. Genera un registro apropiado de datos textuales, abarca un corrector ortográfico [Arruego] y resuelve el tratamiento que se aplica a los números.

[1] AYUDACOD, INE. http://www.ine.es/EX_INICIOAYUDACOD. Instituto Nacional de Estadística (INE), España

[2] hCod Health Codification. http://www.thera-clic.com/site/Productos/hCod-ES.html
AutoCoder. http://www.indizen.com/soluciones/index.html

- Lematización de las respuestas. Permite reducir los términos a su parte esencial, agrupa en una única forma canónica todos los derivados de un lema —se considera como un mismo vocablo las distintas variantes de un mismo lema. Se han utilizado los Servicios Lingüísticos del Grupo de Estructuras de Datos y Lingüística Computacional (GEDLC)[3] de la Universidad de Las Palmas de Gran Canaria [Santana]. De una palabra se pueden obtener varias formas canónicas; a su vez, una misma forma canónica puede haberse generado a partir de diversas palabras.

Resultado de la lematización de: "velas"
Forma Canónica: "velar" CATEG: verbo tran. pron. int
Forma Canónica: "vela" CATEG: sustantivo fem.
Forma Canónica: "ver" CATEG: verbo tran. pron.
Forma Canónica: "ir" CATEG: verbo pron. intr.

Resultado de la lematización de: "máquinas"
Forma Canónica: "máquina" CATEG: sustantivo fem.
Resultado de la lematización de: "maquinilla"
Forma Canónica: "máquina" CATEG: sustantivo fem.
Forma Canónica: "maquinilla" CATEG: sustantivo fem.
Resultado de la lematización de: "máquina"
Forma Canónica: "máquina" CATEG: sustantivo fem.

- Generación de grupos de palabras relacionadas semánticamente. Se explotarán a la hora de asignar los códigos a los textos. Al finalizar esta etapa, pueden obtenerse grupos semánticos que agrupen términos como: *"micología", "taxonomía", "hidrobiología" o "histología"*.
- Generación de índices de relevancia de los textos respecto a los posibles códigos a asignar. Utiliza la información semántica fruto del proceso anterior. Para la categorización de las repuestas, se le da un valor alto de relevancia respecto a un código a aquellas respuestas en las que los términos compartan grupos semánticos como *"micología", "taxonomía", "hidrobiología" o "histología"*.

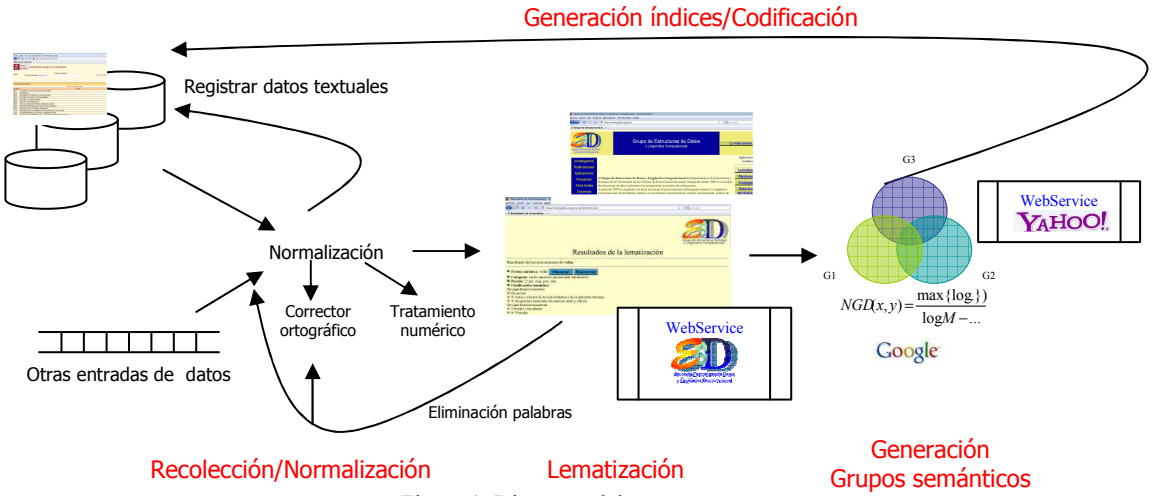

Figura 1. Diagrama del proceso.

2. GENERACIÓN DE GRUPOS SEMÁNTICOS

Los grupos semánticos se construyen a partir de las siguientes hipótesis:

- Los lemas que aparecen en una respuesta están relacionados semánticamente: produce la inclusión directa en un grupo de todos los que figuren en una misma respuesta.
- Si un lema: *A Rsem B y B Rsem C entonces, se acepta A Rsem C*: lleva a incluir en el grupo a todas las palabras "*alcanzables*" desde una dada —semilla del grupo.

1.- Partiendo de la palabra semilla: **"forestal"** y teniendo una respuesta: *"Podador forestal"* → se introduce **"podador"** en el grupo creado
2.- Si **"podador"** aparece en otra respuesta como: *"Podador de árboles "* →se añade **"árboles"** a este mismo grupo

[3] GEDLC Grupo de Estructuras de Datos y Lingüística Computacional. http://www.gedlc.ulpgc.es/. Universidad de Las Palmas de Gran Canaria, España

Aplicar estos criterios origina un conjunto de grupos semánticos que necesita refinarse, lo que se lleva a cabo al seguir los criterios del análisis clúster: se reorganizan los grupos de manera que los resultantes presenten suficiente cohesión interna y marcadas diferencias con los restantes. Se utilizan medidas de similitud semántica en el sentido de indicador de la relación entre términos. La métrica aplicada se basa en el contexto en que se encuentran las palabras. Esta aproximación establece como términos con alta relación semántica aquellos que aparecen en el mismo contexto, frente a otras definiciones que hacen uso de algún espacio semántico como EuroWordNet[4].
En este estudio, el ámbito del contexto de una palabra dada se reduce a la respuesta en que se ha utilizado, lo que produce en general una cantidad de muestras insuficiente de cara a cuantificar la fuerza de la relación semántica entre vocablos. Semejante carencia se subsana en la fase de refinamiento gracias a la Distancia de Google Normalizada (NGD) [Cilibrasi], tal medida tiene por finalidad ponderar cuan cerca están dos términos en el conjunto de documentos que indiza Google; se adopta que tamaño volumen es una muestra robusta que puede tomarse como corpus representativo de la lengua actual.

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

donde

- *f(x), f(y)* nº de páginas que contienen el término *x* e *y respectivamente*
- *f(x, y)* nº de páginas en las que aparecen simultáneamente los términos *x* e *y*.
- válido para M cualquier cota superior del nº de documentos en que aparece la palabra

Estos datos se obtienen utilizando Yahoo como motor de búsqueda, se almacena el número de resultados devueltos en las páginas de España para cada forma canónica o palabra; y también los resultados devueltos para todas las combinaciones de formas canónicas y/o palabras que pertenezcan al mismo grupo. El valor de *M* se obtiene de la cantidad de resultados en el mismo ámbito cuando se solicita la búsqueda de "a".
Se han valorado los motores de búsqueda Google, Yahoo y Bing para obtener los valores utilizados en la expresión para cada par de términos. Google se desechó puesto que limita la cantidad de búsquedas que se pueden realizar. Si bien Bing no presenta este problema, la cantidad de documentos indexados actualmente es mucho menor, lo que hace que la muestra, y por tanto los valores obtenidos, no sea tan fiable. Por último, los resultados obtenidos con Yahoo son similares a los de Google pero con la ventaja de poderse utilizar *Yahoo! Search BOSS*, una nueva plataforma de servicios web que permite el acceso a sus motores de búsqueda sin apenas restricciones, motivo por el cual se seleccionó para el desarrollo de la aplicación.

2.1 Campo de pruebas y resultados

El diseño de la estrategia a seguir se ha experimentado sobre la descripción de forma extensa, en una pregunta abierta, de la ocupación. El concepto de ocupación entraña alguna dificultad práctica para su concreción, y viene definida por el tipo de trabajo realizado y la cualificación. La codificación de la ocupación se hace utilizando la Clasificación Nacional de Ocupaciones de 1994 (C.N.O.-94) a tres dígitos elaborada por el Instituto Nacional de Estadística —de uso obligatorio en el ámbito de la Ley de la Función Estadística Pública a partir de Mayo de 1994. El objetivo de esta clasificación es garantizar el tratamiento uniforme de los datos estadísticos sobre ocupaciones en el ámbito nacional y su comparabilidad comunitaria e internacional.
Se han extraído grupos semánticos a partir de 11140 respuestas sobre ocupación que conforman una muestra de texto libre al que aplicarle el método propuesto. Cada grupo de partida obtenido directamente de ellas está comprendido por conjuntos de palabras heterogéneas que son refinados, disgregándose al finalizar el proceso en subgrupos que constituyen el conjunto de grupos semánticos resultantes de la ejecución del proceso descrito en 2. Se realizaron pruebas con los valores 0.5, 0.6 y 0.7 de la distancia NGD para discriminar la pertenencia al grupo. En general, al usar los límites 0.6, 0.7 no se delimitan claramente conceptos en los grupos, frente al margen de 0.5 que se mostró capaz de generar agrupaciones homogéneas de palabras en torno a un determinado concepto además de colocaciones.
A partir de la proporción entre el máximo rango de entre los subgrupos y el rango del grupo:

$$PRang = \frac{\max_{S \subseteq G}\left(\max_{x \in S, y \in S} NGD(x,y) - \min_{x' \in S, y' \in S} NGD(x',y')\right)}{\max_{x \in G, y \in G} NGD(x,y) - \min_{x' \in G, y' \in G} NGD(x',y')}$$

[4] EuroWordNet. http://www.illc.uva.nl/EuroWordNet/. Universidad de Amsterdam, The Netherlands

se ha evaluado como indicador de la mejora en la cohesión semántica de los grupos resultantes frente a los originales, que se interpreta como un porcentaje de lo que se logra reducir la variabilidad del peor de los subgrupos que surgen a partir de un determinado grupo (Figura 2). Se puede apreciar que la concentración de grupos en los que no aparece mejora se produce entre aquellos en que inicialmente la variabilidad es mínima. Por otra parte, aquellos en los que originalmente la variabilidad es máxima, el peor de los subgrupos originados reduce en la mayor parte de los casos entre un 90% y un 50% su variabilidad.

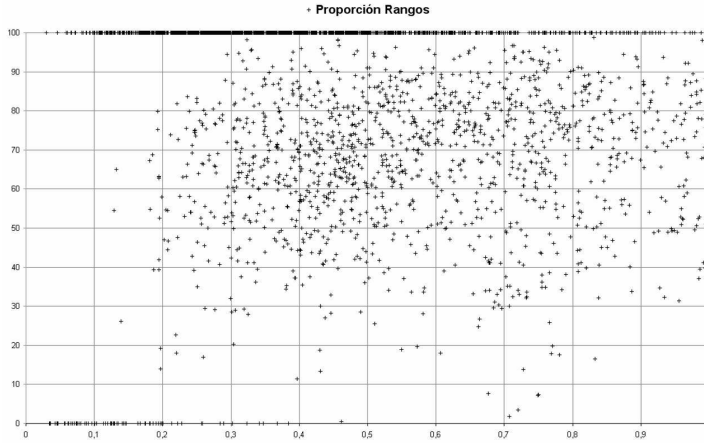

Figura 2. Cambios en la variabilidad entre el grupo original y los subgrupos originados por el refinamiento.

Con objeto de verificar la validez de la clasificación alcanzada, se constata que las palabras incluidas en los grupos semánticos obtenidos aparecen entre las respuestas a las encuestas que comparten un mismo código de ocupación CNO-94. Algunos ejemplos que reflejan esta situación se encuentran en la correspondencia entre el código *7804*[5] y el grupo formado por: *leche, aceite, vegetales, frutas, verduras, margarina, mayonesa,* o el código *2111*[6] y el grupo: *taxonomía, histología, micología, hidrología.*

3. CONCLUSIÓN

El desarrollo de un codificador automático de preguntas de respuesta abierta ha exigido la elaboración de un módulo de generación de grupos semánticos sin necesidad de contar con corpus etiquetados, o cualquier otro recurso semántico. Se ha desarrollado un módulo aplicable a distintos problemas gracias al uso del criterio de similitud semántica basado en la distancia NGD para realizar las agrupaciones en conceptos. Si bien se trata de un proyecto en desarrollo, los resultados obtenidos son adecuados para los datos con los que se ha experimentado, en los que se usa un dominio concreto, pero que abarca un campo semántico amplio.

REFERENCIAS

Arruego J. et al, 2007. Minería de Direcciones Postales. *Actas del V Taller de Minería de Datos y Aprendizaje.* Zaragoza, España, pp. 49-56.

Cilibrasi, R. and Vitányi, P., 2007. The Google Similarity Distance. *In IEEE Transactions on Data and Knowlodge Engineering,* Vol. 19, NO. 3, pp. 370-383

Sánchez Cuadrado, S. et. al. 2009. Extracción Automática de Relaciones Semánticas. *Revista Iberoamericana de Sistemas, Cibernética e Informática,* Vol. 4, NO 2

Santana, O. et al. 2007. Development of Support Services for Linguistic Research over the Internet TIN2004-03988. *Jornadas de Seguimiento de Proyectos en Tecnologías Informáticas*. Madrid, pp. 167-174

[5] Respuestas a las que se les asigna el código 7804: *"deshidratador artesanal, frutas y verduras para conservas", "elaborador de aceites, artesanal", "operario elaboración de productos derivados del aceite mayonesa, margarina, etc", "prensador artesanal, zumos frutas"*

[6] Respuestas a las que se les asigna el código 2111: *"biólogo, hidrobiología", "botánico, ecología", "botánico, genética", "botánico, histología", "botánico, micología", "botánico, taxonomía*