Boletín de Lingüística
ISSN: 0798-9709
vicrag@gmail.com
Universidad Central de Venezuela
Venezuela

García Menier, Everardo
Análisis de textos por computadora
Boletín de Lingüística, vol. XVIII, núm. 25, enero-junio, 2006, pp. 121 -134
Universidad Central de Venezuela
Caracas, Venezuela

NOTAS

ANÁLISIS DE TEXTOS POR COMPUTADORA

Everardo García Menier
Universidad Veracruzana
evgarcia@uv.mx

INTRODUCCIÓN

En la actualidad, la cantidad de información que puede almacenarse de forma electrónica en una computadora es gigantesca. Para darse una idea de esto pensemos que un texto de 15 páginas utiliza alrededor de 100 Kilobytes[1] (Kb) de espacio en disco para almacenarse. Ultimamente, un disco duro promedio posee la capacidad para almacenar 80 Gigabytes (Gb). Existe una unidad intermedia entre los Kb y los Gb, el Megabyte (Mb); Las equivalencias entre estas unidades de medida son las siguientes:

1 Mb	Equivale a	1024 Kb
1 Gb	Equivale a	1024 Mb

de la tabla anterior se tendrá que 1Gb = (1024 x 1024) Kb es decir:
1 Gb = 1,048,576 Kb o sea

que $1Kb = \frac{1}{1{,}048{,}576}\, Gb$. Entonces:

$$100\ Kb = 100\ \frac{1}{1{,}048{,}576}\ Gb = \frac{100}{1{,}048{,}576}\ Gb = 0{,}000095367\ \text{Gb} \approx 0{,}0001\ \text{Gb}$$

Puesto que 100 Kb equivalen aproximadamente a 0,0001 Gb (una diezmilésima parte de un Gb) y que un documento de 15 páginas utiliza alrededor de 100 Kb para almacenarse; entonces 10.000 documentos de 15 páginas utilizarán 10.000 x 0,0001 = 1 Gb. Ahora bien, si tenemos un disco con una capacidad de 80 Mb,[2] éste será capaz de almacenar 10.000 x 80 = 800.000 documentos.

---

1. El Kilobyte es una medida de la capacidad para almacenar información en dispositivos como un disco o la memoria de una computadora.

2. Al momento de escribir este documento, un disco duro de 80 Gb ya se está considerando obsoleto ya que existen discos de 120 o más Gb que son muy accesibles.

"experimento" propuesto en 1950 por el matemático inglés Alan M. Turing (1912-1954) fue una de los disparadores que dio origen a la Inteligencia Artificial.

El experimento consiste en tener dos habitaciones aisladas y conectadas solamente por dos computadoras de tal forma que quienes las usen puedan dialogar entre ellos por medio del teclado y el monitor.[7] Se supone ahora que una persona entra a una de estas habitaciones y se pone a "platicar" con la persona que está en la otra habitación. Si la persona que entró en la primera habitación sale de ella convencido de que sostuvo una conversación con otra persona, pero en la otra habitación sólo se encontraba una computadora que estaba leyendo los mensajes y contestándolos, entonces se diría que la computadora es inteligente. Para decirlo de una forma más correcta, se podría afirmar que el programa en la computadora es un programa inteligente y estaría ubicado en el marco de la Inteligencia Artificial.

El programa inteligente debería contener tres módulos para realizar su tarea: i) un módulo que comprendiera el lenguaje, usado por la persona (lenguaje natural); ii) otro módulo que generara expresiones correctas y entendibles; y iii) un tercer módulo que fuera capaz de "pensar" cuales serían las respuestas que debería dar a la persona. Aquí se marca el nacimiento del Procesamiento de Lenguaje Natural. Sin embargo, en el mundo entero los investigadores de esta área coinciden en que un programa que interprete el lenguaje natural en general es imposible.[8]

Si no se puede hacer que una computadora "comprenda" el lenguaje ¿cómo sería posible entonces analizar un texto? Una de las soluciones sería diseñar programas capaces de analizar un texto a partir de restricciones en el tema tratado, pero éstos deben pertenecer a un dominio específico. Antes de pasar a revisar los enfoques que se han seguido para lograr esto, mencionaremos algunos tipos de análisis de textos que puede realizar una computadora.

2. Análisis de textos por computadora

Hemos planteado la posibilidad de que una computadora realice el análisis de un texto, pero, hasta el momento, no hemos mencionado en qué consiste este análisis. A continuación listaremos los tipos de análisis automático

7. Esta es una versión actualizada del experimento puesto que en aquella época todavía no existían computadoras que pudieran hacer esto.

8. Los más optimistas dicen que es imposible en el corto o mediano plazo.

de textos que una computadora podría realizar. Es importante remarcar el hecho de que ya existe una gran cantidad de programas que realizan este tipo de tareas. Las tareas aquí presentadas no se basan meramente en especulaciones teóricas de lo que se podría lograr con la ayuda de la computadora, sino que constituyen herramientas desarrolladas, probadas, y con un desempeño bastante bueno.

Entre los usos más importantes, se podrían mencionar los siguientes:

- **Recuperación de Información:** Consiste en almacenar una gran cantidad de textos sobre diversos temas para que la computadora seleccione los que esten relacionados con un tema específico. Las búsquedas por Internet serían un ejemplo de esta tarea. El tema se especifica mediante una serie de palabras o frases. El programa, llamado HERRAMIENTA DE BÚSQUEDA,[9] revisa todos los documentos a los que tiene acceso vía Internet y proporciona una lista de aquellos que tratan sobre el tema solicitado. Sin embargo, estas herramientas de búsqueda sólo devuelven los textos que contengan las palabras proporcionadas, por lo que la mayoría de las veces devuelve una colección de cientos de miles de documentos, de los cuales más de la mitad no tienen que ver con el tema buscado. Debe mencionarse que actualmente se están desarrollando herramientas más exactas.
- **Extracción de Información:** Se almacenan los textos en la computadora y se hacen preguntas que la computadora debe responder basándose en los textos analizados. En el programa desarrollado en la Universidad de Utah (Riloff y Lenhert 1994), por ejemplo, los textos proporcionados son notas periodísticas relacionadas con el terrorismo y las preguntas son del tipo *¿Cuántos atentados con bomba se efectuaron entre determinadas fechas?* o *¿Cuántas víctimas de secuestro ha habido en el año?*
- **Clasificación de Textos:** Se proporciona a la computadora un conjunto de textos referentes a diversos temas y el programa debe agruparlos en bloques temáticos. Por ejemplo, agrupa los textos que tratan de deportes o de política o de educación, etc.
- **Autoría de textos:** La computadora recibe una serie de textos del mismo autor y, mediante un análisis, "aprende" el "estilo" del mismo. Después, se le proporciona un texto más y el programa debe

9. *Search engine*

ser capaz de decidir si fue escrito por el mismo autor o por otro diferente y determinar si el "estilo" corresponde o no con el que la computadora aprendió y reconoce.

- **Análisis de tendencias:** Se almacena en la computadora una serie de notas periodísticas acerca de un tema específico y el programa debe generar un informe acerca de las tendencias que el público tiene, con base en las notas procesadas. Un ejemplo es el análisis de la opinión que la gente tiene acerca de un personaje público como podría ser un cantante o un político; basado en la información contenida en los artículos, el programa debe poder detectar si la gente piensa que es un buen cantante o político o si su popularidad se debe sólo a la mercadotecnia o, incluso, si la gente piensa que este cantante o político es una persona amable o es un déspota.
- **Resumen automático de textos:** Se da como entrada a la computadora un texto[10] y el programa debe escribir un resumen del mismo. Debe resaltarse que se busca como resultado un buen resumen, es decir, un texto de longitud significativamente menor a la del texto original y que contenga las ideas y conceptos relevantes del mismo. Un programa de este tipo sería de gran utilidad, por ejemplo, para las personas que padecen afasia.[11]
- **Descubrimiento de conocimiento en bases de datos textuales:** Este tipo de análisis es relativamente nuevo y consiste en que la computadora extrae de los textos almacenados algún conocimiento que no está explícito en el texto o bien puede estar distribuido en varios textos de la colección. Se puede dar a la computadora una gran cantidad de artículos periodísticos que traten sobre diversos temas. El programa, después de haberlas analizado, puede descubrir cosas que no estaban explícitas en ninguna de ellas. El programa puede descubrir, por ejemplo, que la cantidad de accidentes en las carreteras se incrementa durante ciertos meses del año o que los asaltos en una ciudad son más frecuentes en ciertos sitios que en otros.[12] Es muy importante resaltar que esta tarea es llevada a cabo de

10. O bien una serie de textos.

11. A *grosso modo* la afasia es una enfermedad que impide a quien la padece leer textos largos ya que no pueden retener en la memoria el texto completo y olvidan el principio del mismo cuando van leyendo la parte media.

12. Como cerca de las terminales de autobuses o en las salidas de los supermercados.

forma automática por la computadora; el humano no interviene en absoluto en la búsqueda de estas conclusiones.[13] Este tipo de análisis se conoce como MINERÍA DE TEXTOS[14] y es de resaltarse que ya existen varias empresas en el mundo que se dedican a esta tarea.

3. MÉTODOS DE ANÁLISIS

Para lograr que una computadora realice análisis de textos como los mencionados arriba, se han seguido diversos métodos. A continuación se mencionan algunos:

Para tareas de clasificación de textos se ha utilizado un enfoque estadístico que consiste en encontrar, para cada palabra que contiene, su función de distribución de probabilidad.[15] Una vez realizado esto para todos los textos, se comparan estas funciones de distribución y cuando se encuentra que para dos textos estas funciones son muy parecidas, entonces se afirma que ambos se refieren al mismo tema y se agrupan juntos. En caso contrario, los textos son separados, ya que los temas que tratan son diferentes.

Otro enfoque matemático para resolver este problema consiste en convertir a cada texto en un VECTOR,[16] basándose en la frecuencia con la que cada palabra aparece en el mismo. Si pensamos en vectores de dos dimensiones, estaremos hablando de parejas de números que pueden ser fácilmente representadas gráficamente como puntos en un plano cartesiano. Uniendo el origen de este plano con los puntos, obtendremos flechas que forman un ángulo entre ellas. Tomando como base la idea de representar a los vectores como flechas que forman un ángulo entre ellas, en el Álgebra Lineal se extiende, mediante fórmulas, el concepto de ángulo y puede calcularse el ángulo que forman dos vectores en cualquier dimensión. Volviendo al caso de la clasificación de textos, se mide el ángulo entre dos vectores (textos) y, si éste es pequeño, se considera

13. Es obvio que fue un humano quien escribió el programa que hace que la computadora pueda realizar esta tarea, pero ahí termina su intervención.

14. *Text Mining*

15. Para aclarar lo que es una función de distribución de probabilidad, piénsese en la función de distribución Binomial que se estudia en la Estadística Básica. Claro que, hablando de funciones para palabras encontrarlas, resulta mucho más complejo.

16. Podemos visualizar a un vector como una serie ordenada de números; por ejemplo (25, 32, 40, 10) o (42, 38, 21, 34). Por contener cuatro componentes cada uno, se dice que su dimensión es 4.

que los textos tratan sobre temas parecidos y se agrupan juntos; en caso contrario, se considera que tratan de temas diferentes.

En el marco de los eventos llamados *Conferencia de Comprensión de Mensajes*,[17] se proporciona a los sistemas participantes una gran cantidad de textos sobre un tema para que realicen un análisis con el fin de responder a las preguntas. Las respuestas pueden estar en uno de los textos analizados o distribuidas en varios de ellos. Por ejemplo, en el caso del sistema que analiza notas periodísticas relacionadas con el terrorismo, una pregunta podría ser *¿qué grupo perpetró el atentado contra el Diplomático X?* En este caso, la respuesta se encontrará revisando un solo texto que hable de dicho atentado.[18] Pero si se pregunta *¿cuántos atentados realizó el grupo Z en la semana pasada?*, entonces se tendrá que revisar todos los textos que responsabilicen de actos terroristas a este grupo, con el fin de reunir esta información para proporcionar una respuesta.

Se realizó un estudio sobre los métodos que usaron los sistemas en una de estas conferencias. La conclusión fue que existen básicamente dos enfoques y que la mayoría de los sistemas eran híbridos en el sentido de que utilizaban una combinación de ambos, Los enfoques encontrados fueron:

- ❖ El enfoque guiado por sintaxis y
- ❖ El enfoque guiado por semántica.

El primero de ellos consiste en analizar únicamente la estructura sintáctica del texto sin tomar en cuenta la semántica del mismo. Este enfoque presenta algunos problemas:

a) No considera el contexto en que se encuentra la oración analizada, lo que puede conducir a interpretaciones erróneas, como se ejemplifica más adelante,
b) No toma en cuenta el significado de lo que se está analizando y, por lo mismo,
c) No puede análizar más de lo que está escrito.

---

17. *Message Understanding Conference* (MUC)

18. Debe considerarse que este problema puede volverse más complejo, por ejemplo, si hay contradicción en las notas y mencionan cuando menos a dos grupos distintos como los autores del atentado.

Se tratará de aclarar estas limitaciones.

a) Al no tomar en cuenta al contexto, puede incurrirse en interpretaciones erróneas. Supongamos que se está analizando una oración que habla muy bien de la persona X. Si sólo se presta atención a este hecho, puede llegarse a la conclusión de que esta persona es popular y aceptada por la gente. Sin embargo, si el contexto en que está inmersa esta oración es sarcástico, entonces la conclusión a la que se llegaría es diametralmente opuesta a la anterior.
b) Si no se toma en cuenta el significado de la oración que se está analizando, pueden llegar a aceptarse como correctas cosas que carecen de sentido. Por ejemplo, un análisis basado en la sintaxis tomará como correcta tanto la expresión Juanito arrojó una piedra contra el espejo como a su equivalente sintáctico El espejo arrojó una piedra contra Juanito, la cual es inadecuada. Se ha intentado resolver este problema creando gramáticas ad hoc para el análisis por computadora. En el ejemplo de arriba, la situación podría salvarse si se clasificara a los sustantivos en dos grupos: los que pueden realizar acciones (como Juanito) y los que no pueden hacerlo (como el espejo). Así, la segunda expresión se calificaría como semánticamente incorrecta, puesto que una palabra que pertenece a la clase que no puede realizar acciones no suele anteceder a un verbo transitivo. Pero éste es sólo un caso; como él hay muchos, por lo que las gramáticas que se han desarrollado a veces se vuelven extremadamente complejas, con el consecuente consumo de recursos computacionales, lo que hace muy difícil su aplicación.
c) Al restringir el análisis a la "corrección"[19] sintáctica, lo que puede obtenerse es restringido. Retomando el ejemplo de arriba, las únicas preguntas que pueden responderse a partir de la expresión *Juanito arrojó una piedra contra el espejo* son: *¿quién arrojó la piedra? ¿contra quién se arrojó la piedra?* y *¿qué hizo Juanito?* Pero si introdujéramos algo de semántica y algún "conocimiento" adicional, el sistema podría contestar a la pregunta, nada trivial para una computadora: *¿qué le ocurrió al espejo?* Para nosotros como humanos, ésta resulta ser una pregunta trivial con una respuesta

19. *Correctness*

obvia: *El espejo se rompe*; pero, para que una computadora pueda contestar esto, se le debe proporcionar el conocimiento de que el espejo es un vidrio, que el vidrio se rompe al recibir un impacto y que las piedras arrojadas contra un objeto hacen que éste reciba un impacto. Además, se debe dotar a la computadora de un mecanismo que le permita "razonar" con todo este conocimiento como premisa, y llegar a la conclusión de que el espejo se rompe.

d) Otro problema que presenta este enfoque es que en ocasiones aparecen en los textos expresiones sintácticamente incorrectas, que un lector humano es capaz de comprender, pero que no aporta información en el análisis automátizado.

El enfoque guiado por semántica no requiere necesariamente de un análisis sintáctico, sino que centra su atención en el contenido semántico de la expresión analizada; es decir, toma como unidades básicas del análisis expresiones que tienen un significado.

Para seguir este enfoque es necesario dotar a la computadora con "conocimiento" para que sea capaz de "entender" el texto que está analizando. Jacobs *et al*. (1993) consideran que este enfoque es el que mejor se puede implementar en una computadora para aplicaciones prácticas del mundo real.

Una técnica que se ha seguido para la implementación de Sistemas de Análisis de Textos siguiendo este enfoque es la de ir tomando las palabras que componen la expresión hasta obtener una serie de palabras cuya semántica proporcione alguna información. Por ejemplo, si se empieza a analizar una expresión que comienza con la palabra "el", esta palabra no aporta información alguna. Si la segunda palabra es "perro", lo único que podemos obtener es que se está hablando de un perro específico; si la tercera palabra es "negro" sólo sabemos el color de este animal. La siguiente palabra es "mordió"; aunque ahora se conoce la acción realizada por el perro, todavía no se tiene mucha información. La palabra que sigue es "al"; la información que se tiene ahora no difiere de la que ya se tenía. Pero si la otra palabra es "Alcalde", de inmediato se tiene una serie de palabras que proporcionan información relevante, pues ahora sabemos que *el perro negro mordió al Alcalde*. Resulta evidente la dificultad que existe para instruir a una computadora para que encuentre esta clase de expresiones pues ¿qué puede hacerse para saber que las palabras que ha ido tomando ya aportan información relevante y no necesita tomar más palabras?

Podría tomarse como ejemplo un sistema desarrollado en Inglaterra

para analizar las historias clínicas de los pacientes que el hospital elabora al darlos de alta. Este sistema puede dar respuesta a la pregunta anterior. Para esto se elaboraron "plantillas" que se van llenando si el texto analizado coincide con alguna de ellas. La estructura de tales plantillas puede representarse como sigue: [<a/ b/ c> <d/ e> <f> <g/ h>] donde *a, b, c, d, e, f, g, h* son palabras o grupos de palabras. Lo encerrado entre los símbolos [ ] es la estructura que se buscará en el texto, lo que aparece entre los símbolos < > puede constar de una sola palabra como <f> o una serie de palabras de las cuales se puede escoger cualquiera, así <a/ b/ c> significa que en ese sitio puede ir cualquiera de las palabras *a, b* o *c*.

Una de tales plantillas podría ser [<el / la> <prueba / examen / análisis/ estudio> <hace sospechar/ indica/ comprueba> <la> <presencia/ ausencia> <de> <estafilococo/ neumococo/ salmonella/ amiba/ gérmenes patógenos> <en> <vías respiratorias/ vías urinarias/ amígdalas/ estómago>]. El procedimiento que realiza la computadora es revisar las palabras en orden y ver si la secuencia se ajusta a esta plantilla. Así, por ejemplo, si se encuentra la secuencia: *el análisis comprueba la ausencia de gérmenes patógenos en vías respiratorias* será aceptada por el sistema.[20] Sin embargo, el caso de la secuencia *el estudio hace sospechar que el grado de avance de la enfermedad es menor al que se pensaba* es diferente. Aunque la primera parte de la secuencia *el estudio hace sospechar* concuerda con la plantilla, el resto ya no. Esto implica que esta plantilla no es la adecuada para analizar esta expresión en particular. Hay que crear una nueva plantilla que contemple este tipo de casos. Al terminar el análisis del texto, se tendrá un conjunto de secuencias con significado. El manejo de estas secuencias permitirá que se obtenga la información requerida por el especialista (en este caso, el médico).

Debe notarse que este análisis puede ser realizado de manera autónoma por la computadora, si ya se tiene el conjunto de plantillas a que debe sujetarse. Otro punto que merece resaltarse es la gran cantidad de expresiones que tiene el lenguaje médico, lo cual hace que escribir las plantillas sea una tarea sumamente difícil (Mikheev 1996).

Otro ejemplo de un sistema basado en semántica es el de un clasificador de notas periodísticas. En este caso se proporciona a la computadora un conjunto de noticias que tratan acerca de una institución educativa. La tarea que la computadora debe realizar es clasificar estas noticias en dos grupos: los que

20. Debe notarse que cada palabra que aparece en la secuencia aparece también en la plantilla y en ese orden. Una vez llena esta plantilla, se habrá obtenido una frase que aporta información.

causan una buena impresión de la institución a los lectores y los que causan una mala impresión. Para realizarla, se hizo un listado de palabras que, para los fines de una institución educativa, causan una buena impresión (por ejemplo, *buenos_egresados*,[21] *infraestructura*, *biblioteca_completa*) y otro listado con palabras que causan una mala impresión (por ejemplo, *malos_egresados*, *ausentismo_de_profesores*, *huelga*). A continuación se listan palabras que indican presencia como *tiene*, *posee*, *se_da*, *hay*, etc. y palabras que denotan ausencia como falta, carece, escasez; también se consideró la palabra *no* que cambia el sentido de la expresión.

Una vez hecho lo anterior se dieron las siguientes reglas:

a. Presencia de cosas buenas equivale a una buena impresión;
b. Ausencia de cosas buenas equivale a una mala impresión;
c. Presencia de cosas malas equivale a una mala impresión y;
d. Ausencia de cosas malas equivale a una buena impresión.

Para realizar el análisis de una noticia se procede así: Si se encuentra en el texto la secuencia *esta institución posee una biblioteca completa* se considera que causará una buena impresión (regla *a*). Por otro lado, al hallar la secuencia *en esta institución se da un alto índice de ausentismo por parte de los profesores* (regla *c*) se considera que causa una mala impresión. Al final se realiza un conteo de las secuencias y de la impresión que causan. Si el número de las expresiones que causan una buena impresión es significativamente mayor al número de las que causan una mala, se considera que la nota está causando una buena impresión en los lectores. En caso contrario, se considera que la impresión causada es mala. Si ambos números son muy parecidos no se extrae ninguna conclusión.

De esta manera se pueden ir clasificando las notas en buenas y malas para la institución y así se obtendrá una idea de cómo la considera el público.[22]

Existe otro enfoque surgido de las teorías desarrolladas por el matemático Zellig Harris (1909-1992), quien fuera profesor de Noam Chomsky, famoso entre otras cosas por sus trabajos sobre ligüística. Este enfoque se basa en el concepto de SUBLENGUAJE.

---

21. Aunque esta y algunas de las que se mencionan abajo no son palabras aisladas se consideraron como si fuera una sola "palabra", de ahí el uso del símbolo "_" para unirlas.

22. Este sistema se describe en García Menier (1998).

Zellig Harris (1982) propuso una teoría de los sublenguajes que explica por qué es posible procesar el lenguaje en textos especializados pertenecientes a dominios específicos tales como los encontrados en genética y medicina.

> De acuerdo con Harris, los lenguajes de Dominios técnicos tienen estructura y regularidad, que pueden ser observadas examinando los léxicos de los Dominios y que pueden delinearse de tal forma que la estructura puede especificarse de una forma apropiada para la computadora. Mientras que la teoría general de la gramática Inglesa [o de otro idioma] especifica primordialmente sólo estructuras sintácticamente bien formadas, la teoría de la gramática de los sublenguajes de Harris también incorpora información semántica específica del Dominio y las relaciones para delinear un lenguaje que es más informativo que el Inglés porque refleja el objeto de estudio y las relaciones del Dominio así como la estructura sintáctica (Friedman *et al*. 2002: 223).

La teoría de los sublenguajes tiene una sólida base matemática. Actualmente este enfoque está siendo utilizado con éxito por varios investigadores de diversas partes del mundo. Por ejemplo, en la Universidad de Columbia se han construido sistemas para analizar textos pertenecientes al dominio de la medicina. En uno de ellos se proporciona a la computadora una serie de interpretaciones de estudios radiológicos; el sistema re-escribe estas interpretaciones pero en un lenguaje estandarizado, lo cual las hace susceptibles de un tratamiento por computadora. Este enfoque tiene aplicaciones para el análisis de textos de muy diversos dominios.

4. Conclusión

Como se ha visto, el análisis de textos por computadora es una herramienta que tiene mucha utilidad en la solución de problemas en diversas áreas. Esperamos que algunos lectores se sientan atraídos por esta mezcla de elementos lingüísticos y computacionales para beneficio de ambas disciplinas. Estamos abiertos a recibir opiniones, sugerencias, críticas que puedan ayudarnos a enriquecer el trabajo. Todas estas sugerencias son bienvenidas en la dirección electrónica del autor: evgarcia@uv.mx.

REFERENCIAS BIBLIOGRÁFICAS

Friedman, Carol; Pauline Kra y Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35. 222–235.

García Menier, Everardo. 1998. Un sistema para la clasificación de notas periodísticas. En Pedro Galicia (ed.), *Memorias del Simposium Internacional de Computación La Computación: Investigación. desarrollo y aplicaciones*, 197-204. México, DF: Instituto Politécnico Nacional.

Harris, Zellig. 1982. Discourse and sublanguage. En Richard Kittredge y John Lehrberger (eds.), *Sublanguages: Studies on language in restricted semantic domains*, 231-236. Berlín: Walter de Gruyter.

Jacobs, Paul; George Krupka; Lisa Rau; Michael Mauldin; Teruko Mitamura; Tsuyoshi Kitani; Ira Sider y Lois Childs. 1993. [En línea]. GE-CMU: *Description of the Shogun System used for the Fifth Message Understanding Conference (MUC-5)*. Disponible en www.cs.mu.oz.au/acl/M/M93/M93-1011.pdf [Consulta: 18 de enero 2005].

Mikheev, Andrei. 1996. Domain knowledge for natural language processing. *Research Papers HCRC RP-70*, 1-33. Edinburgh: Human Communication Research Centre. Language Technology Group. University of Edinburgh.

Riloff, Ellen y Wendy Lenhert. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on information systems* 2, 3. 296-333.

El cálculo anterior muestra la gran cantidad de textos que se puede almacenar en un solo disco. Resulta prácticamente impensable realizar un análisis de ochocientos mil documentos en forma manual. Para ello, se requiere forzosamente del uso de la computadora. Es más, lo que se pretende es que sea la propia computadora, sin la intervención del ser humano, la que realice el análisis.

Sin embargo, la computadora no es capaz de diferenciar los datos que tiene almacenados; por ejemplo, resultará indistinguible un archivo que contiene las ventas diarias de un supermercado de uno que contenga un artículo periodístico. Para diferenciarlos, se debe alimentar a la computadora con el programa adecuado.

Cuando los datos almacenados en la computadora tienen una estructura bien definida, puede obtenerse una gran cantidad de información útil. Supóngase que se tiene una agencia de viajes y el problema a resolver consiste en encontrar los posibles vuelos, escalas, transbordos, etc. que un cliente podría realizar para llegar al destino que desea y, posiblemente, hacer escalas en ciertas ciudades. Para resolver este problema, puede crearse una Base de Datos[3] en la cual se almacenarían, por ejemplo, todas las ciudades de destino, los puntos de partida, las compañías aéreas que tienen vuelos entre estas ciudades, los costos de los boletos, los enlaces con otras compañías, etc. Esta información está perfectamente estructurada y con ella se puede resolver, de manera sencilla, el problema planteado. Si un cliente desea salir de la ciudad *A* y llegar a la ciudad *B*, haciendo una escala en la ciudad *C*, como toda esta información ya se encuentra almacenada, la computadora tiene que explorar las posibilidades para tal viaje, calculando para cada una de ellas su costo, el tiempo necesario para el recorrido total, los tiempos de espera entre transbordos, etc. y con esto se tendrá un planteamiento para hacer al cliente. Así, éste podrá escoger la combinación que más le agrade.[4]

Pero ¿qué sucede si los datos almacenados en la computadora son textos y el problema consiste en realizar un análisis de los mismos? En este caso, la tarea se vuelve extremadamente compleja, puesto que la computadora

---

3. Entenderemos por Base de Datos un conjunto de archivos en los que cada uno de ellos contiene información acerca de un asunto particular perfectamente ordenada y, además, estos archivos están relacionados entre si, de una forma bien estructurada para poder compartir la información que poseen.

4. En realidad este problema no es tan sencillo, puesto que se vuelve tremendamente complejo cuando la cantidad de ciudades es muy grande. Sin embargo, pensamos que es un buen ejemplo del uso que puede darse a la información que se encuentra bien estructurada.

no "entiende" lo que significan estos datos. Por ejemplo, podemos escribir *El gato duerme plácidamente a la sombra del árbol* o *El caballo dictó una conferencia que arrancó una gran ovación al público* o *incluso verde flores automóvil teléfono*. Como se puede observar, la primera de estas afirmaciones es perfectamente comprensible y posible; la segunda, aunque es comprensible, resulta absurda; y la tercera no es sino un conjunto de palabras que no tiene ningún sentido. Sin embargo, al almacenarlas en la computadora, ésta no será capaz de distinguir las que tienen significado de las que no lo tienen. Para ella estos datos tienen tanto sentido y validez como un conjunto de letras generado aleatoriamente o una tabla de tipos de cambio de monedas.[5] Por tanto, es necesario encontrar el tratamiento que una computadora debe dar a los textos almacenados que se van a analizar.

El propósito del presente trabajo es dar una breve introducción al análisis de textos por computadora, las aplicaciones que pueden darse a este tipo de análisis, así como los métodos que se han utilizado para contender con este problema.

1. Un poco de historia

El problema de que una computadora analice, de manera autónoma, un texto o un conjunto de ellos cae dentro de la Inteligencia Artificial,[6] una rama de la computación que puede definirse como sigue: si un ente artificial ejecuta una conducta que puede ser considerada inteligente entonces se está haciendo Inteligencia Artificial. En el caso que nos ocupa, el ente artificial es la computadora y la conducta, obviamente inteligente, que presenta es la de analizar un texto. Puede realizar tareas como, por ejemplo, extraer conclusiones acerca de un texto, responder a preguntas con base en el análisis realizado, dividir un conjunto de textos en grupos de manera tal que cada uno de ellos trate del mismo tema, etc.

Una disciplina que dio origen al análisis de textos por computadora es el Procesamiento del Lenguaje Natural. En el marco de esta disciplina se pretendía que una computadora interpretara cualquier texto con el fin de encontrar la semántica asociada al conjunto de palabras que lo conforman. Un

5. En este último caso sería muy sencillo, utilizando estos datos, convertir bolívares a pesos peruanos o yenes a quetzales.

6. La Inteligencia Artificial se ha definido de otras formas, pero la definición presentada aquí es la que mejor se ajusta para el presente trabajo.

"experimento" propuesto en 1950 por el matemático inglés Alan M. Turing (1912-1954) fue una de los disparadores que dio origen a la Inteligencia Artificial.

El experimento consiste en tener dos habitaciones aisladas y conectadas solamente por dos computadoras de tal forma que quienes las usen puedan dialogar entre ellos por medio del teclado y el monitor.[7] Se supone ahora que una persona entra a una de estas habitaciones y se pone a "platicar" con la persona que está en la otra habitación. Si la persona que entró en la primera habitación sale de ella convencido de que sostuvo una conversación con otra persona, pero en la otra habitación sólo se encontraba una computadora que estaba leyendo los mensajes y contestándolos, entonces se diría que la computadora es inteligente. Para decirlo de una forma más correcta, se podría afirmar que el programa en la computadora es un programa inteligente y estaría ubicado en el marco de la Inteligencia Artificial.

El programa inteligente debería contener tres módulos para realizar su tarea: i) un módulo que comprendiera el lenguaje, usado por la persona (lenguaje natural); ii) otro módulo que generara expresiones correctas y entendibles; y iii) un tercer módulo que fuera capaz de "pensar" cuales serían las respuestas que debería dar a la persona. Aquí se marca el nacimiento del Procesamiento de Lenguaje Natural. Sin embargo, en el mundo entero los investigadores de esta área coinciden en que un programa que interprete el lenguaje natural en general es imposible.[8]

Si no se puede hacer que una computadora "comprenda" el lenguaje ¿cómo sería posible entonces analizar un texto? Una de las soluciones sería diseñar programas capaces de analizar un texto a partir de restricciones en el tema tratado, pero éstos deben pertenecer a un dominio específico. Antes de pasar a revisar los enfoques que se han seguido para lograr esto, mencionaremos algunos tipos de análisis de textos que puede realizar una computadora.

2. Análisis de textos por computadora

Hemos planteado la posibilidad de que una computadora realice el análisis de un texto, pero, hasta el momento, no hemos mencionado en qué consiste este análisis. A continuación listaremos los tipos de análisis automático

7. Esta es una versión actualizada del experimento puesto que en aquella época todavía no existían computadoras que pudieran hacer esto.

8. Los más optimistas dicen que es imposible en el corto o mediano plazo.

de textos que una computadora podría realizar. Es importante remarcar el hecho de que ya existe una gran cantidad de programas que realizan este tipo de tareas. Las tareas aquí presentadas no se basan meramente en especulaciones teóricas de lo que se podría lograr con la ayuda de la computadora, sino que constituyen herramientas desarrolladas, probadas, y con un desempeño bastante bueno.

Entre los usos más importantes, se podrían mencionar los siguientes:

- **Recuperación de Información:** Consiste en almacenar una gran cantidad de textos sobre diversos temas para que la computadora seleccione los que esten relacionados con un tema específico. Las búsquedas por Internet serían un ejemplo de esta tarea. El tema se especifica mediante una serie de palabras o frases. El programa, llamado HERRAMIENTA DE BÚSQUEDA,[9] revisa todos los documentos a los que tiene acceso vía Internet y proporciona una lista de aquellos que tratan sobre el tema solicitado. Sin embargo, estas herramientas de búsqueda sólo devuelven los textos que contengan las palabras proporcionadas, por lo que la mayoría de las veces devuelve una colección de cientos de miles de documentos, de los cuales más de la mitad no tienen que ver con el tema buscado. Debe mencionarse que actualmente se están desarrollando herramientas más exactas.
- **Extracción de Información:** Se almacenan los textos en la computadora y se hacen preguntas que la computadora debe responder basándose en los textos analizados. En el programa desarrollado en la Universidad de Utah (Riloff y Lenhert 1994), por ejemplo, los textos proporcionados son notas periodísticas relacionadas con el terrorismo y las preguntas son del tipo *¿Cuántos atentados con bomba se efectuaron entre determinadas fechas?* o *¿Cuántas víctimas de secuestro ha habido en el año?*
- **Clasificación de Textos:** Se proporciona a la computadora un conjunto de textos referentes a diversos temas y el programa debe agruparlos en bloques temáticos. Por ejemplo, agrupa los textos que tratan de deportes o de política o de educación, etc.
- **Autoría de textos:** La computadora recibe una serie de textos del mismo autor y, mediante un análisis, "aprende" el "estilo" del mismo. Después, se le proporciona un texto más y el programa debe

9. *Search engine*

ser capaz de decidir si fue escrito por el mismo autor o por otro diferente y determinar si el "estilo" corresponde o no con el que la computadora aprendió y reconoce.

- **Análisis de tendencias:** Se almacena en la computadora una serie de notas periodísticas acerca de un tema específico y el programa debe generar un informe acerca de las tendencias que el público tiene, con base en las notas procesadas. Un ejemplo es el análisis de la opinión que la gente tiene acerca de un personaje público como podría ser un cantante o un político; basado en la información contenida en los artículos, el programa debe poder detectar si la gente piensa que es un buen cantante o político o si su popularidad se debe sólo a la mercadotecnia o, incluso, si la gente piensa que este cantante o político es una persona amable o es un déspota.
- **Resumen automático de textos:** Se da como entrada a la computadora un texto[10] y el programa debe escribir un resumen del mismo. Debe resaltarse que se busca como resultado un buen resumen, es decir, un texto de longitud significativamente menor a la del texto original y que contenga las ideas y conceptos relevantes del mismo. Un programa de este tipo sería de gran utilidad, por ejemplo, para las personas que padecen afasia.[11]
- **Descubrimiento de conocimiento en bases de datos textuales:** Este tipo de análisis es relativamente nuevo y consiste en que la computadora extrae de los textos almacenados algún conocimiento que no está explícito en el texto o bien puede estar distribuido en varios textos de la colección. Se puede dar a la computadora una gran cantidad de artículos periodísticos que traten sobre diversos temas. El programa, después de haberlas analizado, puede descubrir cosas que no estaban explícitas en ninguna de ellas. El programa puede descubrir, por ejemplo, que la cantidad de accidentes en las carreteras se incrementa durante ciertos meses del año o que los asaltos en una ciudad son más frecuentes en ciertos sitios que en otros.[12] Es muy importante resaltar que esta tarea es llevada a cabo de

10. O bien una serie de textos.

11. A *grosso modo* la afasia es una enfermedad que impide a quien la padece leer textos largos ya que no pueden retener en la memoria el texto completo y olvidan el principio del mismo cuando van leyendo la parte media.

12. Como cerca de las terminales de autobuses o en las salidas de los supermercados.

"experimento" propuesto en 1950 por el matemático inglés Alan M. Turing (1912-1954) fue una de los disparadores que dio origen a la Inteligencia Artificial.

El experimento consiste en tener dos habitaciones aisladas y conectadas solamente por dos computadoras de tal forma que quienes las usen puedan dialogar entre ellos por medio del teclado y el monitor.[7] Se supone ahora que una persona entra a una de estas habitaciones y se pone a "platicar" con la persona que está en la otra habitación. Si la persona que entró en la primera habitación sale de ella convencido de que sostuvo una conversación con otra persona, pero en la otra habitación sólo se encontraba una computadora que estaba leyendo los mensajes y contestándolos, entonces se diría que la computadora es inteligente. Para decirlo de una forma más correcta, se podría afirmar que el programa en la computadora es un programa inteligente y estaría ubicado en el marco de la Inteligencia Artificial.

El programa inteligente debería contener tres módulos para realizar su tarea: i) un módulo que comprendiera el lenguaje, usado por la persona (lenguaje natural); ii) otro módulo que generara expresiones correctas y entendibles; y iii) un tercer módulo que fuera capaz de "pensar" cuales serían las respuestas que debería dar a la persona. Aquí se marca el nacimiento del Procesamiento de Lenguaje Natural. Sin embargo, en el mundo entero los investigadores de esta área coinciden en que un programa que interprete el lenguaje natural en general es imposible.[8]

Si no se puede hacer que una computadora "comprenda" el lenguaje ¿cómo sería posible entonces analizar un texto? Una de las soluciones sería diseñar programas capaces de analizar un texto a partir de restricciones en el tema tratado, pero éstos deben pertenecer a un dominio específico. Antes de pasar a revisar los enfoques que se han seguido para lograr esto, mencionaremos algunos tipos de análisis de textos que puede realizar una computadora.

2. Análisis de textos por computadora

Hemos planteado la posibilidad de que una computadora realice el análisis de un texto, pero, hasta el momento, no hemos mencionado en qué consiste este análisis. A continuación listaremos los tipos de análisis automático

7. Esta es una versión actualizada del experimento puesto que en aquella época todavía no existían computadoras que pudieran hacer esto.

8. Los más optimistas dicen que es imposible en el corto o mediano plazo.

de textos que una computadora podría realizar. Es importante remarcar el hecho de que ya existe una gran cantidad de programas que realizan este tipo de tareas. Las tareas aquí presentadas no se basan meramente en especulaciones teóricas de lo que se podría lograr con la ayuda de la computadora, sino que constituyen herramientas desarrolladas, probadas, y con un desempeño bastante bueno.

Entre los usos más importantes, se podrían mencionar los siguientes:

- **Recuperación de Información:** Consiste en almacenar una gran cantidad de textos sobre diversos temas para que la computadora seleccione los que esten relacionados con un tema específico. Las búsquedas por Internet serían un ejemplo de esta tarea. El tema se especifica mediante una serie de palabras o frases. El programa, llamado HERRAMIENTA DE BÚSQUEDA,[9] revisa todos los documentos a los que tiene acceso vía Internet y proporciona una lista de aquellos que tratan sobre el tema solicitado. Sin embargo, estas herramientas de búsqueda sólo devuelven los textos que contengan las palabras proporcionadas, por lo que la mayoría de las veces devuelve una colección de cientos de miles de documentos, de los cuales más de la mitad no tienen que ver con el tema buscado. Debe mencionarse que actualmente se están desarrollando herramientas más exactas.
- **Extracción de Información:** Se almacenan los textos en la computadora y se hacen preguntas que la computadora debe responder basándose en los textos analizados. En el programa desarrollado en la Universidad de Utah (Riloff y Lenhert 1994), por ejemplo, los textos proporcionados son notas periodísticas relacionadas con el terrorismo y las preguntas son del tipo *¿Cuántos atentados con bomba se efectuaron entre determinadas fechas?* o *¿Cuántas víctimas de secuestro ha habido en el año?*
- **Clasificación de Textos:** Se proporciona a la computadora un conjunto de textos referentes a diversos temas y el programa debe agruparlos en bloques temáticos. Por ejemplo, agrupa los textos que tratan de deportes o de política o de educación, etc.
- **Autoría de textos:** La computadora recibe una serie de textos del mismo autor y, mediante un análisis, "aprende" el "estilo" del mismo. Después, se le proporciona un texto más y el programa debe

9. *Search engine*

ser capaz de decidir si fue escrito por el mismo autor o por otro diferente y determinar si el "estilo" corresponde o no con el que la computadora aprendió y reconoce.

- **Análisis de tendencias:** Se almacena en la computadora una serie de notas periodísticas acerca de un tema específico y el programa debe generar un informe acerca de las tendencias que el público tiene, con base en las notas procesadas. Un ejemplo es el análisis de la opinión que la gente tiene acerca de un personaje público como podría ser un cantante o un político; basado en la información contenida en los artículos, el programa debe poder detectar si la gente piensa que es un buen cantante o político o si su popularidad se debe sólo a la mercadotecnia o, incluso, si la gente piensa que este cantante o político es una persona amable o es un déspota.
- **Resumen automático de textos:** Se da como entrada a la computadora un texto[10] y el programa debe escribir un resumen del mismo. Debe resaltarse que se busca como resultado un buen resumen, es decir, un texto de longitud significativamente menor a la del texto original y que contenga las ideas y conceptos relevantes del mismo. Un programa de este tipo sería de gran utilidad, por ejemplo, para las personas que padecen afasia.[11]
- **Descubrimiento de conocimiento en bases de datos textuales:** Este tipo de análisis es relativamente nuevo y consiste en que la computadora extrae de los textos almacenados algún conocimiento que no está explícito en el texto o bien puede estar distribuido en varios textos de la colección. Se puede dar a la computadora una gran cantidad de artículos periodísticos que traten sobre diversos temas. El programa, después de haberlas analizado, puede descubrir cosas que no estaban explícitas en ninguna de ellas. El programa puede descubrir, por ejemplo, que la cantidad de accidentes en las carreteras se incrementa durante ciertos meses del año o que los asaltos en una ciudad son más frecuentes en ciertos sitios que en otros.[12] Es muy importante resaltar que esta tarea es llevada a cabo de

10. O bien una serie de textos.

11. A *grosso modo* la afasia es una enfermedad que impide a quien la padece leer textos largos ya que no pueden retener en la memoria el texto completo y olvidan el principio del mismo cuando van leyendo la parte media.

12. Como cerca de las terminales de autobuses o en las salidas de los supermercados.

forma automática por la computadora; el humano no interviene en absoluto en la búsqueda de estas conclusiones.[13] Este tipo de análisis se conoce como MINERÍA DE TEXTOS[14] y es de resaltarse que ya existen varias empresas en el mundo que se dedican a esta tarea.

3. MÉTODOS DE ANÁLISIS

Para lograr que una computadora realice análisis de textos como los mencionados arriba, se han seguido diversos métodos. A continuación se mencionan algunos:

Para tareas de clasificación de textos se ha utilizado un enfoque estadístico que consiste en encontrar, para cada palabra que contiene, su función de distribución de probabilidad.[15] Una vez realizado esto para todos los textos, se comparan estas funciones de distribución y cuando se encuentra que para dos textos estas funciones son muy parecidas, entonces se afirma que ambos se refieren al mismo tema y se agrupan juntos. En caso contrario, los textos son separados, ya que los temas que tratan son diferentes.

Otro enfoque matemático para resolver este problema consiste en convertir a cada texto en un VECTOR,[16] basándose en la frecuencia con la que cada palabra aparece en el mismo. Si pensamos en vectores de dos dimensiones, estaremos hablando de parejas de números que pueden ser fácilmente representadas gráficamente como puntos en un plano cartesiano. Uniendo el origen de este plano con los puntos, obtendremos flechas que forman un ángulo entre ellas. Tomando como base la idea de representar a los vectores como flechas que forman un ángulo entre ellas, en el Álgebra Lineal se extiende, mediante fórmulas, el concepto de ángulo y puede calcularse el ángulo que forman dos vectores en cualquier dimensión. Volviendo al caso de la clasificación de textos, se mide el ángulo entre dos vectores (textos) y, si éste es pequeño, se considera

13. Es obvio que fue un humano quien escribió el programa que hace que la computadora pueda realizar esta tarea, pero ahí termina su intervención.

14. *Text Mining*

15. Para aclarar lo que es una función de distribución de probabilidad, piénsese en la función de distribución Binomial que se estudia en la Estadística Básica. Claro que, hablando de funciones para palabras encontrarlas, resulta mucho más complejo.

16. Podemos visualizar a un vector como una serie ordenada de números; por ejemplo (25, 32, 40, 10) o (42, 38, 21, 34). Por contener cuatro componentes cada uno, se dice que su dimensión es 4.

que los textos tratan sobre temas parecidos y se agrupan juntos; en caso contrario, se considera que tratan de temas diferentes.

En el marco de los eventos llamados *Conferencia de Comprensión de Mensajes*,[17] se proporciona a los sistemas participantes una gran cantidad de textos sobre un tema para que realicen un análisis con el fin de responder a las preguntas. Las respuestas pueden estar en uno de los textos analizados o distribuidas en varios de ellos. Por ejemplo, en el caso del sistema que analiza notas periodísticas relacionadas con el terrorismo, una pregunta podría ser *¿qué grupo perpetró el atentado contra el Diplomático X?* En este caso, la respuesta se encontrará revisando un solo texto que hable de dicho atentado.[18] Pero si se pregunta *¿cuántos atentados realizó el grupo Z en la semana pasada?*, entonces se tendrá que revisar todos los textos que responsabilicen de actos terroristas a este grupo, con el fin de reunir esta información para proporcionar una respuesta.

Se realizó un estudio sobre los métodos que usaron los sistemas en una de estas conferencias. La conclusión fue que existen básicamente dos enfoques y que la mayoría de los sistemas eran híbridos en el sentido de que utilizaban una combinación de ambos, Los enfoques encontrados fueron:

- ❖ El enfoque guiado por sintaxis y
- ❖ El enfoque guiado por semántica.

El primero de ellos consiste en analizar únicamente la estructura sintáctica del texto sin tomar en cuenta la semántica del mismo. Este enfoque presenta algunos problemas:

a) No considera el contexto en que se encuentra la oración analizada, lo que puede conducir a interpretaciones erróneas, como se ejemplifica más adelante,
b) No toma en cuenta el significado de lo que se está analizando y, por lo mismo,
c) No puede análizar más de lo que está escrito.

---

17. *Message Understanding Conference* (MUC)

18. Debe considerarse que este problema puede volverse más complejo, por ejemplo, si hay contradicción en las notas y mencionan cuando menos a dos grupos distintos como los autores del atentado.

Se tratará de aclarar estas limitaciones.

a) Al no tomar en cuenta al contexto, puede incurrirse en interpretaciones erróneas. Supongamos que se está analizando una oración que habla muy bien de la persona X. Si sólo se presta atención a este hecho, puede llegarse a la conclusión de que esta persona es popular y aceptada por la gente. Sin embargo, si el contexto en que está inmersa esta oración es sarcástico, entonces la conclusión a la que se llegaría es diametralmente opuesta a la anterior.
b) Si no se toma en cuenta el significado de la oración que se está analizando, pueden llegar a aceptarse como correctas cosas que carecen de sentido. Por ejemplo, un análisis basado en la sintaxis tomará como correcta tanto la expresión Juanito arrojó una piedra contra el espejo como a su equivalente sintáctico El espejo arrojó una piedra contra Juanito, la cual es inadecuada. Se ha intentado resolver este problema creando gramáticas ad hoc para el análisis por computadora. En el ejemplo de arriba, la situación podría salvarse si se clasificara a los sustantivos en dos grupos: los que pueden realizar acciones (como Juanito) y los que no pueden hacerlo (como el espejo). Así, la segunda expresión se calificaría como semánticamente incorrecta, puesto que una palabra que pertenece a la clase que no puede realizar acciones no suele anteceder a un verbo transitivo. Pero éste es sólo un caso; como él hay muchos, por lo que las gramáticas que se han desarrollado a veces se vuelven extremadamente complejas, con el consecuente consumo de recursos computacionales, lo que hace muy difícil su aplicación.
c) Al restringir el análisis a la "corrección"[19] sintáctica, lo que puede obtenerse es restringido. Retomando el ejemplo de arriba, las únicas preguntas que pueden responderse a partir de la expresión *Juanito arrojó una piedra contra el espejo* son: *¿quién arrojó la piedra? ¿contra quién se arrojó la piedra?* y *¿qué hizo Juanito?* Pero si introdujéramos algo de semántica y algún "conocimiento" adicional, el sistema podría contestar a la pregunta, nada trivial para una computadora: *¿qué le ocurrió al espejo?* Para nosotros como humanos, ésta resulta ser una pregunta trivial con una respuesta

19. *Correctness*

obvia: *El espejo se rompe*; pero, para que una computadora pueda contestar esto, se le debe proporcionar el conocimiento de que el espejo es un vidrio, que el vidrio se rompe al recibir un impacto y que las piedras arrojadas contra un objeto hacen que éste reciba un impacto. Además, se debe dotar a la computadora de un mecanismo que le permita "razonar" con todo este conocimiento como premisa, y llegar a la conclusión de que el espejo se rompe.

d) Otro problema que presenta este enfoque es que en ocasiones aparecen en los textos expresiones sintácticamente incorrectas, que un lector humano es capaz de comprender, pero que no aporta información en el análisis automátizado.

El enfoque guiado por semántica no requiere necesariamente de un análisis sintáctico, sino que centra su atención en el contenido semántico de la expresión analizada; es decir, toma como unidades básicas del análisis expresiones que tienen un significado.

Para seguir este enfoque es necesario dotar a la computadora con "conocimiento" para que sea capaz de "entender" el texto que está analizando. Jacobs *et al*. (1993) consideran que este enfoque es el que mejor se puede implementar en una computadora para aplicaciones prácticas del mundo real.

Una técnica que se ha seguido para la implementación de Sistemas de Análisis de Textos siguiendo este enfoque es la de ir tomando las palabras que componen la expresión hasta obtener una serie de palabras cuya semántica proporcione alguna información. Por ejemplo, si se empieza a analizar una expresión que comienza con la palabra "el", esta palabra no aporta información alguna. Si la segunda palabra es "perro", lo único que podemos obtener es que se está hablando de un perro específico; si la tercera palabra es "negro" sólo sabemos el color de este animal. La siguiente palabra es "mordió"; aunque ahora se conoce la acción realizada por el perro, todavía no se tiene mucha información. La palabra que sigue es "al"; la información que se tiene ahora no difiere de la que ya se tenía. Pero si la otra palabra es "Alcalde", de inmediato se tiene una serie de palabras que proporcionan información relevante, pues ahora sabemos que *el perro negro mordió al Alcalde*. Resulta evidente la dificultad que existe para instruir a una computadora para que encuentre esta clase de expresiones pues ¿qué puede hacerse para saber que las palabras que ha ido tomando ya aportan información relevante y no necesita tomar más palabras?

Podría tomarse como ejemplo un sistema desarrollado en Inglaterra

para analizar las historias clínicas de los pacientes que el hospital elabora al darlos de alta. Este sistema puede dar respuesta a la pregunta anterior. Para esto se elaboraron "plantillas" que se van llenando si el texto analizado coincide con alguna de ellas. La estructura de tales plantillas puede representarse como sigue: [<a/ b/ c> <d/ e> <f> <g/ h>] donde *a, b, c, d, e, f, g, h* son palabras o grupos de palabras. Lo encerrado entre los símbolos [ ] es la estructura que se buscará en el texto, lo que aparece entre los símbolos < > puede constar de una sola palabra como <f> o una serie de palabras de las cuales se puede escoger cualquiera, así <a/ b/ c> significa que en ese sitio puede ir cualquiera de las palabras *a, b* o *c*.

Una de tales plantillas podría ser [<el / la> <prueba / examen / análisis/ estudio> <hace sospechar/ indica/ comprueba> <la> <presencia/ ausencia> <de> <estafilococo/ neumococo/ salmonella/ amiba/ gérmenes patógenos> <en> <vías respiratorias/ vías urinarias/ amígdalas/ estómago>]. El procedimiento que realiza la computadora es revisar las palabras en orden y ver si la secuencia se ajusta a esta plantilla. Así, por ejemplo, si se encuentra la secuencia: *el análisis comprueba la ausencia de gérmenes patógenos en vías respiratorias* será aceptada por el sistema.[20] Sin embargo, el caso de la secuencia *el estudio hace sospechar que el grado de avance de la enfermedad es menor al que se pensaba* es diferente. Aunque la primera parte de la secuencia *el estudio hace sospechar* concuerda con la plantilla, el resto ya no. Esto implica que esta plantilla no es la adecuada para analizar esta expresión en particular. Hay que crear una nueva plantilla que contemple este tipo de casos. Al terminar el análisis del texto, se tendrá un conjunto de secuencias con significado. El manejo de estas secuencias permitirá que se obtenga la información requerida por el especialista (en este caso, el médico).

Debe notarse que este análisis puede ser realizado de manera autónoma por la computadora, si ya se tiene el conjunto de plantillas a que debe sujetarse. Otro punto que merece resaltarse es la gran cantidad de expresiones que tiene el lenguaje médico, lo cual hace que escribir las plantillas sea una tarea sumamente difícil (Mikheev 1996).

Otro ejemplo de un sistema basado en semántica es el de un clasificador de notas periodísticas. En este caso se proporciona a la computadora un conjunto de noticias que tratan acerca de una institución educativa. La tarea que la computadora debe realizar es clasificar estas noticias en dos grupos: los que

20. Debe notarse que cada palabra que aparece en la secuencia aparece también en la plantilla y en ese orden. Una vez llena esta plantilla, se habrá obtenido una frase que aporta información.

causan una buena impresión de la institución a los lectores y los que causan una mala impresión. Para realizarla, se hizo un listado de palabras que, para los fines de una institución educativa, causan una buena impresión (por ejemplo, *buenos_egresados*,[21] *infraestructura*, *biblioteca_completa*) y otro listado con palabras que causan una mala impresión (por ejemplo, *malos_egresados*, *ausentismo_de_profesores*, *huelga*). A continuación se listan palabras que indican presencia como *tiene*, *posee*, *se_da*, *hay*, etc. y palabras que denotan ausencia como falta, carece, escasez; también se consideró la palabra *no* que cambia el sentido de la expresión.

Una vez hecho lo anterior se dieron las siguientes reglas:

a. Presencia de cosas buenas equivale a una buena impresión;
b. Ausencia de cosas buenas equivale a una mala impresión;
c. Presencia de cosas malas equivale a una mala impresión y;
d. Ausencia de cosas malas equivale a una buena impresión.

Para realizar el análisis de una noticia se procede así: Si se encuentra en el texto la secuencia *esta institución posee una biblioteca completa* se considera que causará una buena impresión (regla *a*). Por otro lado, al hallar la secuencia *en esta institución se da un alto índice de ausentismo por parte de los profesores* (regla *c*) se considera que causa una mala impresión. Al final se realiza un conteo de las secuencias y de la impresión que causan. Si el número de las expresiones que causan una buena impresión es significativamente mayor al número de las que causan una mala, se considera que la nota está causando una buena impresión en los lectores. En caso contrario, se considera que la impresión causada es mala. Si ambos números son muy parecidos no se extrae ninguna conclusión.

De esta manera se pueden ir clasificando las notas en buenas y malas para la institución y así se obtendrá una idea de cómo la considera el público.[22]

Existe otro enfoque surgido de las teorías desarrolladas por el matemático Zellig Harris (1909-1992), quien fuera profesor de Noam Chomsky, famoso entre otras cosas por sus trabajos sobre ligüística. Este enfoque se basa en el concepto de SUBLENGUAJE.

---

21. Aunque esta y algunas de las que se mencionan abajo no son palabras aisladas se consideraron como si fuera una sola "palabra", de ahí el uso del símbolo "_" para unirlas.

22. Este sistema se describe en García Menier (1998).

Zellig Harris (1982) propuso una teoría de los sublenguajes que explica por qué es posible procesar el lenguaje en textos especializados pertenecientes a dominios específicos tales como los encontrados en genética y medicina.

> De acuerdo con Harris, los lenguajes de Dominios técnicos tienen estructura y regularidad, que pueden ser observadas examinando los léxicos de los Dominios y que pueden delinearse de tal forma que la estructura puede especificarse de una forma apropiada para la computadora. Mientras que la teoría general de la gramática Inglesa [o de otro idioma] especifica primordialmente sólo estructuras sintácticamente bien formadas, la teoría de la gramática de los sublenguajes de Harris también incorpora información semántica específica del Dominio y las relaciones para delinear un lenguaje que es más informativo que el Inglés porque refleja el objeto de estudio y las relaciones del Dominio así como la estructura sintáctica (Friedman *et al*. 2002: 223).

La teoría de los sublenguajes tiene una sólida base matemática. Actualmente este enfoque está siendo utilizado con éxito por varios investigadores de diversas partes del mundo. Por ejemplo, en la Universidad de Columbia se han construido sistemas para analizar textos pertenecientes al dominio de la medicina. En uno de ellos se proporciona a la computadora una serie de interpretaciones de estudios radiológicos; el sistema re-escribe estas interpretaciones pero en un lenguaje estandarizado, lo cual las hace susceptibles de un tratamiento por computadora. Este enfoque tiene aplicaciones para el análisis de textos de muy diversos dominios.

4. Conclusión

Como se ha visto, el análisis de textos por computadora es una herramienta que tiene mucha utilidad en la solución de problemas en diversas áreas. Esperamos que algunos lectores se sientan atraídos por esta mezcla de elementos lingüísticos y computacionales para beneficio de ambas disciplinas. Estamos abiertos a recibir opiniones, sugerencias, críticas que puedan ayudarnos a enriquecer el trabajo. Todas estas sugerencias son bienvenidas en la dirección electrónica del autor: evgarcia@uv.mx.

REFERENCIAS BIBLIOGRÁFICAS

Friedman, Carol; Pauline Kra y Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35. 222–235.

García Menier, Everardo. 1998. Un sistema para la clasificación de notas periodísticas. En Pedro Galicia (ed.), *Memorias del Simposium Internacional de Computación La Computación: Investigación. desarrollo y aplicaciones*, 197-204. México, DF: Instituto Politécnico Nacional.

Harris, Zellig. 1982. Discourse and sublanguage. En Richard Kittredge y John Lehrberger (eds.), *Sublanguages: Studies on language in restricted semantic domains*, 231-236. Berlín: Walter de Gruyter.

Jacobs, Paul; George Krupka; Lisa Rau; Michael Mauldin; Teruko Mitamura; Tsuyoshi Kitani; Ira Sider y Lois Childs. 1993. [En línea]. GE-CMU: *Description of the Shogun System used for the Fifth Message Understanding Conference (MUC-5)*. Disponible en www.cs.mu.oz.au/acl/M/M93/M93-1011.pdf [Consulta: 18 de enero 2005].

Mikheev, Andrei. 1996. Domain knowledge for natural language processing. *Research Papers HCRC RP-70*, 1-33. Edinburgh: Human Communication Research Centre. Language Technology Group. University of Edinburgh.

Riloff, Ellen y Wendy Lenhert. 1994. Information extraction as a basis for high-precision text classification. *ACM Transactions on information systems* 2, 3. 296-333.