Citación: Ordoñez, S.S. y Gelbukh, A. (2010). Representación computacional del lenguaje natural escrito. En: *Ingeniería, Vol. 15, No. 1, pág. 6 -21.*

Artículo de Revisión

Representación computacional del lenguaje natural escrito

Computational representation of written natural language

Sonia Ordoñez Salinas
Universidad Distrital
Francisco José de Caldas
Facultad de Ingeniería
sordonez@udistrital.edu.co

Alexander Gelbukh
Laboratorio de Lenguaje Natural y Procesamiento de Texto del Centro de Investigación en Computación del Instituto Politécnico Nacional, México
www.gelbukh.com

Resumen

Cuando el ser humano lee o escucha una palabra, inmediatamente la relaciona con un concepto. Esto es posible gracias a la acumulación de información y a la posibilidad de filtrar, procesar y relacionar dicha información. Para la máquina, una expresión escrita en el lenguaje natural es una cadena de bits que no aporta información por sí sola. Un computador interpreta esta cadena de bits, modelando el proceso que tiene lugar en la mente humana, estructurando y relacionado la cadena con información previamente almacenada. En el proceso, así como al momento de describir los resultados, el texto es representado por estructuras formales que permiten el procesamiento automático, la interpretación y la comparación de la información. Este artículo presenta una descripción detallada de estas estructuras.

Palabras claves: Procesamiento de lenguaje natural, estructuras computacionales.

Abstract

When humans read, or hear, words, they immediately relate them to a concept. This is possible due to the information already stored in the brain and also to human's ability to select, process, and associate such information with words. However, for a computer, natural language text is only a sequence of bits that does not convey any meaning on its own, unless properly processed. A computer interprets this bit sequence by modeling the processing that takes place in human minds, namely structuring and linking the text with previously stored information. During this process, as well as when describing its results, the text is represented using various formal structures that permit automatic processing, interpretation, and comparison of information. In this paper, we present a detailed description of these structures.

Key words: Natural language processing, computational structures.

1. Introducción

La mayor parte de la información que posee la humanidad se encuentra almacenada en forma de lenguaje natural. La gran necesidad de los usuarios de esta información es gestionarla: almacenarla, consultarla, entenderla y actualizarla.

Para el ser humano una frase expresada en lenguaje natural leída, escuchada, hablada o escrita adquiere inmediatamente un significado. Cada palabra o conjunto de palabras, se asocia con un conjunto de experiencias e imágenes que al ser relacionadas y filtradas trae a la mente exactamente el significado correcto. El hombre identifica de quien se habla, la acción que realiza, el tiempo, los sinónimos, el dominio y todas las particularidades del lenguaje natural. Para una máquina computacional, un texto escrito en lenguaje natural, corresponde a una cadena de bits (símbolos) sin significado alguno. Para que una máquina pueda *entender* su significado, se debe recurrir a técnicas propias del procesamiento del lenguaje natural (PLN) y de la lingüística computacional (LC) y a artificios computacionales como el de estructurar la información de tal forma que esta pueda ser gestionada y relacionada con algún significado. Significado que puede buscarse usualmente en elementos especializados para tal fin, tales como ontologías, tesauros, bases de conocimiento, entre otros (ver Figura No 1).

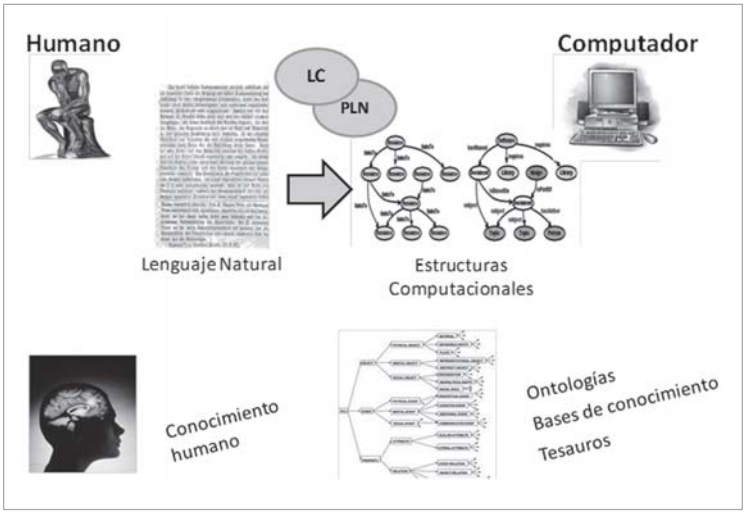

Figura 1. Procesamiento del lenguaje natural por el humano vs. la máquina (fuente: autores)

El proceso de estructurar el texto escrito en lenguaje natural o de representar la cadena de símbolos en una estructura computacional como un vector, un árbol, una pila, un modelo relacional, entre otras, permite que se pueda gestionar la cadena a través de algoritmos computacionales. Para transformar de forma automática el texto a la estructura escogida se utilizan las partes de dicha cadena, es decir las palabras y sus interrelaciones y de acuerdo a la estructura escogida, es deseable que, por un lado, dicha estructura permita incluir la mayor cantidad de información —entendida información como aquella que aporta al significado de la cadena—, y por el otro, que el proceso de transformar la cadena a esta estructura sea factible de ser automatizado.

En términos generales se podría afirmar que el PLN, es una disciplina que con apoyo de la LC se encarga de proveer soluciones para la interpretación y gestión del lenguaje natural. Dichas disciplinas se soportan en técnicas y métodos propios de la estadística, las matemáticas, la lingüística, la inteligencia artificial, entre otras. Lo correspondiente a la generación de herramientas y aplicativos que incluyen la gestión del lenguaje natural se puede enmarcar dentro de lo que se conoce como ingeniería lingüística.

En la actualidad ya no es solo el deseo de que las máquinas computacionales puedan establecer una comunicación utilizando el lenguaje natural, tal y como lo hacen las personas sino que se ha convertido en una exigencia. Es así que responder a un sinnúmero de necesidades actuales como la utilización de la información almacenada a través de la Web o de aplicativos ya sea para buscar información específica, patrones de comportamiento o predicciones han hecho que cada vez se busquen y optimicen alternativas para que de alguna manera las máquinas interpreten el contenido de los documentos digitales expresados en lenguaje natural. Consecuentemente, entre las aplicaciones más comunes en las áreas en mención son la recuperación de información, la respuesta automática a preguntas, la traducción automática y la clasificación de documentos.

En este artículo se presenta una revisión del primer aspecto mencionado, es decir de las estructuras computacionales utilizadas para el procesamiento del lenguaje natural escrito. El artículo está organizado como sigue. En la sección 2 se presenta brevemente la información general sobre dichas estructuras. En la sección 3 se presentan las estructuras básicas, las cuales no representan la semántica detallada. En la sección 4 se presentan estructuras más avanzadas. En la sección 5 presenta detalles sobre las estructuras conceptuales que tienen un estándar claro para su construcción. Finalmente la sección 6 concluye el artículo.

2. Estructuras computacionales en el PLN

Una gran variedad de formalismos estructurados han sido propuestos para representar los textos en el lenguaje natural. Estos formalismos van desde las representaciones más simples como las vectoriales hasta las más complejas como los lenguajes estructurados, estructuras conceptuales y formalismos matemáticos como los grafos. Los formalismos más simples, en el sentido de que pueden capturar menos cantidad de elementos que conduzcan a la interpretación del significado o semántica y su procesamiento, incluida la transformación de la cadena de caracteres a dicha estructura requiere de una algoritmia. Por su parte los formalismos más complejos, si bien permiten incluir más elementos que pueden contribuir a la interpretación del significado inmerso dentro de la cadena, el procesamiento —la transformación de lenguaje natural al formalismo y viceversa— es más complicado y, lo más importante, mucho menos confiable con la tecnología actual. En la sección 4 se amplia y detallan estas estructuras y se justifica del porqué se hacen tales afirmaciones.

En las siguientes secciones se presenta una revisión de las estructuras computacionales, empezando por las más básicas y continuando con más avanzadas.

3. Estructuras básicas

Dentro de esta categoría se clasifican aquellas estructuras que se usan cuando no se requiere mayor nivel de detalle, resultados aceptables y bajo costo computacional. Tanto el proceso de convertir el texto en una estructura, como el de gestionar dicha estructura, requiere de una algoritmia básica y por ende un bajo costo computacional. Es de anotar que estas estructuras también son la base para las estructuras más avanzadas.

3.1 Modelo de espacio vectorial

Teniendo en cuenta el conjunto de palabras que hacen parte de una sentencia, una representación muy clásica es la propuesta por Salton [1], autor del trabajo pionero en representar los documentos como un vector de frecuencias, donde cada entrada (coordenada) del vector corresponde a una determinada palabra dentro del documento y su valor es la frecuencia de aparición de dicha palabra. Se puede decir que la mayoría de trabajos de recuperación de información y clasificación utilizan dicha representación, como en [2], [3] y [4].

Trabajar con las palabras que pertenecen a un lenguaje conlleva al manejo de un gran número de variables y por ende el procesamiento de alta dimensionalidad. El procesar el lenguaje a través de alguna estructura no elimina, por sí solo, la problemática de la dimensionalidad, por lo que aparecen trabajos que estudian el comportamiento de las palabras con análisis estadístico [5] y los que permiten asimilar las palabras a sus raíces (*stemming*) como el presentado en [6] o el algoritmo propuesto por Porter, llamado *Porter stemmer* [7].

Otra representación basada también en vectores, es aquella donde cada entrada del vector significa la presencia o ausencia de una palabra en el documento, es decir que el documento se representa como un vector de entradas binarias [8].

La representación de documentos con la estructura vectorial, puede darse a través de tuplas como un conjunto de parejas (término, peso), donde el peso corresponde a un valor dado de acuerdo a la parte del documento donde aparece la palabra —por ejemplo, un título tendrá mayor peso que un subtítulo.

La colección de palabras puede ser extraída de todo el documento o de alguna parte específica como de los títulos o el resumen: por ejemplo, en [9] las palabras se extraen únicamente de los enlaces de un hipertexto, para así cumplir con un trabajo de clasificación.

Con esta representación se pueden utilizar diferentes técnicas en las tareas propias del procesamiento del lenguaje natural. Por ejemplo, en los trabajos de recuperación de información y de indexación se utilizan métodos de semántica latente que buscan visualizar las relaciones ocultas que existen entre las palabras a través de la aplicación de vectores y valores propios [10]. Estas técnicas se aplican sobre vectores de frecuencia o sobre vectores cuyo contenido corresponde a valores binarios [11].

3.2 Listas

Las colecciones de palabras representadas con estructuras básicas no solo se han tenido en cuenta de forma vectorial sino también como listas. Den-tro de esta clasificación se pueden referenciar las estructuras utilizadas para indexar documentos a través de los índices invertidos. Un índice invertido es una lista de términos acompañados por los números o referencias de los documentos en los cuales aparecen dichos términos. El contenido de dicha lista puede tener algunas variaciones como por ejemplo que a demás de contener el documento, contenga las posiciones donde se encuentra la palabra dentro del documento o la presentada en [12], donde cada documento se representa por medio de una lista de tuplas (palabra, valor).

3.3 Grafos

Existen varios métodos para representar los documentos como grafos. En [13], se clasifican los métodos en: estándares, simples, distancia *n*, distancia *n*-simple, frecuencia absoluta y frecuencia relativa. Cada método se basa en examinar los términos en cada documento y sus adyacencias. Al igual que en otros métodos, los términos se extraen del documento y se realiza un pre-procesamiento, que generalmente consiste en eliminar las palabras que no aportan significado a los documentos (*stop words*) para así tratar de disminuir la dimensionalidad. A continuación brevemente se expondrá en qué consiste cada una de las representaciones.

- **Estándar.** Cada palabra corresponde a la etiqueta de un nodo y si una palabra *a* inmediatamente precede a una palabra *b* en una sección *s*, entonces existe una arista que comienza en *a* y termina en *b* etiquetada con *s*. En este caso se tiene en cuenta la puntuación y las secciones propias de un documento como el título o resumen, entre otros.

  Simple. A diferencia de la estándar, no se etiquetan las aristas con las secciones y no se tienen en cuenta todas las secciones sino aquellas que sean más visibles.
- **Distancia *n*.** Se buscan las *n* palabras siguientes a partir de un término dado, y las aristas se etiquetan con la distancia al punto inicial. El parámetro *n* es definido por el usuario.
- **Distancia simple.** Es similar a la anterior, con la diferencia de que las aristas no son etiquetadas y lo único que se sabe es que la distancia entre dos nodos conectados es menor que *n*.
- **Frecuencia absoluta.** Es similar a la representación simple, pero cada nodo y arista son etiquetados con una medida de frecuencia. Para un nodo esto indica cuántas veces los términos aparecen en el documento; para las aristas esto significa el número de veces que dos términos conectados aparecen en el orden específico. Bajo esta representación, el tamaño del grafo se define como la suma de las frecuencias de los nodos más la suma de las frecuencias de las aristas.
- **Frecuencia relativa.** Es similar a la frecuencia absoluta con la diferencia de que las frecuencias son normalizadas entre 0 y 1. Los nodos se normalizan por el valor máximo de frecuencia de los nodos y las aristas similarmente por el máximo valor de frecuencia en las aristas.

Varias propuestas utilizan la representación con grafos. Por ejemplo, en [6] se representa un documento por un grafo que tiene en cuenta la frecuencia de ocurrencia de las palabras. En [14] se incluyen varias formas de representar los documentos a través de los grafos. Además, en [6] se propone una metodología para la construcción de grafos: a partir de análisis del dominio se presentan objetos o entidades que son agrupados dentro de clases o tipos.

3.4 Estructuras estadísticas

Con base a la teoría de la información se han realizado investigaciones sobre el comportamiento de las palabras y la información que más aporta a un documento.

El primer modelo estadístico del lenguaje fue propuesto por Claude Shannon [39]. De acuerdo a la teoría de la información, el lenguaje se modela como una fuente estadística de información. La estadística se ha constituido en una herramienta fundamental para el análisis de lenguaje natural. De igual forma se han propuesto estructuras que incluyen funciones probabilísticas para representar el contenido de un texto. Dentro de estas estructuras están, por ejemplo, los modelos de Markov, las gramáticas probabilísticas, los analizadores probabilísticos [22]. En general, cualquier estructura (un vector, un grafo o lista, entre otros) puede ser marcada con probabilidades o funciones de probabilidad.

En los modelos probabilísticos, se construyen distribuciones de los documentos que pertenecen a una clase. En el caso de recuperación de información, por ejemplo, se asume una distribución para los documentos relevantes y otra para los no relevantes. Se establecen suposiciones sobre las distribuciones, como independencia, y se aplican procedimientos como el clasificador de Bayes simplificado (*naïve Bayes*) [22]. En los trabajos como [18], la representación de los documentos se logra a través de funciones probabilísticas, y se asume que un documento fue generado a partir de una función de densidad. En dicha investigación, se utiliza un método de agrupamiento (*clustering*) discriminativo distribucional para extractar las características relevantes de los documentos y así representar los documentos como una distribución de probabilidad.

En otros trabajos se mezclan algunas de las técnicas previamente ya expuestas con modelos del lenguaje. En [19], por ejemplo, se presenta un modelo probabilístico para representar el grafo de un documento, con la esperanza de que en el marco de la recuperación de información, un modelo de grafo pueda generar o estimar un grafo de la consulta.

Los modelos estadísticos diferentes a los modelos probabilísticos incluyen las dependencias directas que se presentan por la proximidad o adyacencia entre palabras como en los bigramas y en general los *n*-gramas ([19], [21]).

4. Estructuras avanzadas

Para este documento se asumen como estructuras avanzadas aquellas que permiten incluir elementos propios de la lingüística. Entendida la lingüística como el estudio de la estructura de las lenguas naturales [52]. La ventaja de incluir más información propia del lenguaje (mayor significado) se torna en desventaja a la hora de su procesamiento. La

tarea para una máquina de *entender* que elemento debe ubicar, donde lo debe ubicar y con quien lo debe relacionar, no es tan trivial como se quisiera. El proceso tanto de transformar de manera automática el texto a la estructura en cuestión como el de gestionar la estructura puede llegar a requerir de métodos heurísticos propios de la inteligencia artificial, el aprendizaje estadístico, entre otros, que por su fundamentación teórica basada generalmente en supuestos estadísticos, arrastran un margen de inexactitud.

Estas afirmaciones se hacen en virtud a que algunas de estas estructuras permiten incluir elementos propios de la sintáctica y la semántica, entre otros, por lo que exige la ubicación y la relación de elementos propios de la interpretación del lenguaje. Si se habla de elementos sintácticos, se requiere de la ubicación, las interrelaciones y las dependencias del verbo, el sustantivo, el predicado, entre otros. Si se habla de elementos semánticos, el objeto, el actor, el recipiente, la herramienta, sus relaciones y dependencias, entre otros.

El proceso de automatización de los anteriores elementos, exige de la interpretación del lenguaje con todos los fenómenos propios, como por ejemplo, que una palabra puede referirse a un sujeto que previamente o posteriormente se ha hecho referencia (anáforas y catáforas), que dos palabras diferentes significan lo mismo (sinonimia), que una palabra significa diferente dependiendo del contexto (homonimia), que un sustantivo puede ejercer la función de adjetivo, que una palabra o conjunto de palabras hacen referencia a un nombre propio (entidad nombrada).

Dentro de esta categoría se pueden ubicar las representaciones a través de las gramáticas y a la lógica, así como los lenguajes orientados a objetos y aquellos usados para la creación de repositorios de datos como el Lenguaje de Definición de Datos (*DDL*, por sus siglas en inglés: *Data Definition Language*).

4.1 Representación gramatical

Una forma que se podría utilizar para representar un documento es a través de la descripción morfológica y sintáctica. La descripción morfológica establece la clase gramatical de cada una de las palabras del texto. Dicha clase se define de acuerdo a la función gramatical en el proceso que se denomina etiquetamiento de categorías gramaticales (*part of speech tagging*).

El análisis sintáctico se encarga de analizar la sentencia teniendo en cuenta tanto la función que ejerce cada palabra así como las relaciones existentes entre las mismas. Para dicho análisis se construyen estructuras sintácticas mediante las técnicas de constituyentes, de dependencias o de enlaces, entre otras.

La técnica de análisis de constituyentes consiste en construir un árbol a partir de la teoría de gramáticas generativas expuestas por Noam Chomsky [53] y de un proceso iterativo de segmentación y clasificación gramatical de la oración, hasta que las partes constituyentes sean indivisibles.

Desde el punto de vista de la teoría de las dependencias [36] se puede construir una jerarquía —esquematizada en un árbol— de acuerdo al papel que ejercen las palabras dentro de la frase como la cabeza o raíz del árbol, las subordinadas y rectoras [37]. Las

relaciones se marcan con flechas y varias palabras pueden depender de una sola, aunque cada palabra excepto la raíz depende de exactamente de otra palabra.

Las gramáticas de enlace (*link grammars*) introducidas en [40] se definen como un conjunto de palabras que requieren de un enlace. Una sucesión de palabras equivale a una frase del lenguaje natural si existe una forma de dibujar los enlaces entre las palabras que la conforman. Dichos enlaces no se cruzan y satisfacen los requerimientos de cada palabra.

4.2 Lógica de primer orden y otros métodos basados en lógica

En la lógica de primer orden (*FOL* por sus siglas en inglés: *First Order Logic*), con pocos símbolos básicos se pueden representar elementos del mundo real y a través de predicados se pueden establecer las relaciones entre los elementos. Es así que usando elementos, proposiciones y operadores simples se puede representar el texto [15].

La lógica del primer orden es un sistema deductivo formal usado en las matemáticas, filosofía, lingüística e informática. Se conoce también como cálculo de predicados de primer orden (*FOPC* por sus siglas en inglés: *First Order Predicate Calculus*), el más bajo cálculo de predicados, el lenguaje lógico de primer orden o lógica de predicados.

Con esta técnica se pueden representar texto y por ende documentos. Por ejemplo, en [16] un documento se presenta como una sentencia lógica proposicional de la forma $d = (\textit{recuperación} \wedge \textit{información}) \vee (\textit{recuperación} \wedge \textit{datos})$.

Siguiendo la representación a través de la lógica, se encuentran investigaciones como [17], donde a través de la lógica difusa se representa el documento para implementar posteriormente una forma de recuperación de documentos.

4.3 Lenguaje de representación de marcos

Los lenguajes de tipo FRL (por sus siglas en inglés: *Frame Representation Language*) clasificados bajo esta categoría, se definen como metalenguajes basados en el concepto de marco (*frame*), orientado al reconocimiento y descripción de objetos y clases. La idea principal del razonamiento basado en estos lenguajes es la simplificación, ya que la unidad primaria de organización es el marco. Un marco tiene un nombre y puede tener varios atributos. Cada atributo tiene a su vez un nombre y un valor. Los diferentes tipos de valores pueden ser de una amplia variedad de acuerdo al sistema de marcos. Los valores más comunes son las cadenas y los símbolos [41].

En los FRLs, la herencia es el mecanismo de inferencia central basado en la organización jerárquica. Muchos sistemas que se basan en este lenguaje, permiten la herencia múltiple, como los lenguajes de programación y en particular los orientados a objetos. Los FRLs, contrario a las ontologías que buscan representar el conocimiento en detalle [15], solo se ocupan de representar el conocimiento como objetos. Sin embargo, algunas implementaciones y lenguajes basados en FRLs, pueden asumir un lenguaje ontológico. Dentro de las implementaciones basadas en FRLs destacan los siguientes:

Protégé. Protégé [42] es una plataforma de desarrollo de ontologías definida en marcos y bajo el estándar del lenguaje ontológico para la Web (*OWL*: *Web Ontology Language*),

desarrollada en Java. Las ontologías definidas pueden ser exportadas en una amplia variedad de formatos.

Capa de inferencia ontológica. La Capa de Inferencia Ontológica o lenguaje de intercambio ontológico (*OIL*: *Ontology Inference Layer* o *Ontology Interchange Language*) [43] se define como una estructura ontológica para la semántica en la Web.

Lógica de marcos. La lógica de marcos (*Frame logic* o *F-logic*) [44] es un lenguaje ontológico que permite representar el conocimiento. Su estructura está basada tanto en el FRL como en la orientación a objetos.

Sistema para representación de conocimiento. El sistema para representación de conocimiento (*Knowledge Representation One* o *KL-ONE*) [45] es un lenguaje muy similar al mismo FRL (aunque los marcos en este lenguaje se llaman conceptos) con la diferencia de que se permite el manejo de subclase, superclase y múltiple herencia.

Lógica de descripciones. La lógica de descripciones (*DL*: *Descriptions Logics*) [46] corresponde a una familia de lenguajes definidos para representar el conocimiento. Su nombre se refiere por un lado a la descripción de conceptos para describir dominios y por otro a la semántica basada en lógica de predicados de primer orden. Las DLs fueron diseñadas como una extensión de los FRLs y las redes semánticas con el fin de fortalecer la parte semántica formal.

La lógica descriptiva no solamente representa el conocimiento como objetos sino justifica el proceso con un razonamiento estrictamente lógico basado en conceptos, roles y relaciones. Los conceptos corresponden a clases de elementos y son tratados como subconjuntos del universo. Las relaciones corresponden a vínculos entre elementos y son tratados como relaciones binarias.

5. Estructuras conceptua-les estandarizadas

Las estructuras conceptuales para la representación de conocimiento son modelos (artefactos) que representan una realidad percibida. Con base en técnicas como la inteligencia artificial, estos modelos además de representar el conocimiento, pueden lograr reconstruir conocimiento a través de métodos como la inferencia [23].

Dentro de esta clasificación se encuentran las redes semánticas, los grafos conceptuales, el formato de intercambio de conocimiento (*KIF*: *Knowledge Interchange Format*) [47], el lenguaje de descripción de recursos (*RDF*: *Resource Description Framework*) [48] del consorcio *World Wide Web Consortium* (W3C) y los diferentes lenguajes ontológicos para la Web (*OWL*: *Web Ontology Language*) propuestos por el W3C [24].

Otra estructura conceptual que se puede mencionar, es la Lógica Común (*CL*: *Common Logic*) que no hace parte del grupo W3C pero se posiciona al lado de RDF y de OWL. CL es un marco (*framework*) para una familia de lenguajes basados en lógica. La lógica de primer orden tiene como objetivo el intercambio y la transmisión de información. El objetivo de este *framework* es proveer la sintaxis y semántica abstracta de las sintácticas

concretas o dialectos. El CL preserva el modelo teórico de primer orden, pero cuenta con características muy particulares como la de permitir construcciones de orden superior tales como cuantificaciones sobre clases o relaciones.

5.1 Redes semánticas

Las redes semánticas aparecen a partir de trabajos lingüísticos presentados en el año 1968, y de ahí en adelante los diferentes aportes llevaron a que a finales de los años 70, se puedan observar dos tendencias: por un lado, las redes estructuradas y los sistemas de representación del conocimiento y por el otro, las multiredes orientadas a las ciencias cognitivas.

De forma general se pude definir una red semántica como un modelo matemático que consiste de una estructura conceptual formada por un conjunto de conceptos y relaciones cognitivas entre éstos. Son representadas por un grafo generalizado donde los conceptos corresponden a los nodos y las relaciones entre los conceptos corresponden a los arcos [25] y, desde el punto de vista semántico, como una colección de las diferentes relaciones que los conceptos tienen entre si [26]. Generalmente, la construcción de una frase se logra con la ayuda de los analizadores sintácticos. Sin embargo, esto es aún un problema para aquellos idiomas que no cuentan con un orden estricto de palabras, como el español [27]. Las redes semánticas se catalogan en tres categorías [28]:

Red de relaciones *es-un*. Se construye desde los conceptos más generalizados hasta más específicos, que se representan de forma jerárquica. Dado que la filosofía de estas redes gira en torno a la herencia y la explicación de los conceptos mediante otros conceptos, son generalmente complejas. Entre las investigaciones que utilicen una red semántica del tipo red de relaciones *es-un* (*is-a* en inglés) se puede mencionar, por ejemplo [29], donde se construye una red en forma de árbol; sus atributos y palabras se seleccionan a través del método de entropía condicional.

Red de marcos. Este tipo de red se representa a través de estructuras de datos llamadas marcos. Cada marco se corresponde a una clase o a una instancia. Las clases describen los conceptos mediante un conjunto de propiedades y los marcos son las instancias de las clases. Ellos describen objetos concretos y heredan propiedades de los marcos clase. Los atributos y valores se esquematizan a través de ranuras (*slots*).

Grafos conceptuales. Durante los años 60 la representación semántica basada en grafos fue popular tanto a nivel teórica como en la lingüística computacional. Esta representación se conoce como grafo conceptual. Esta estructura, propuesta por Sowa [26], está basada en los grafos existenciales de Pierce [32].

El estándar para los grafos conceptuales especifica la sintaxis, la semántica y la representación de cadenas de caracteres en el formato de intercambio de grafos conceptuales (*CGIF*: *Conceptual graph interchange form*). CGIF ha sido diseñado para intercambio de información entre los sistemas que hacen parte del estándar "Formato para el modelamiento de esquemas conceptuales" (*CSMF*: *Conceptual Schema Modeling Facilities*) especificado por el estándar ISO/IEC 14481. El estándar de los GC provee una guía para implementar sistemas que usan los grafos conceptuales en la representación interna

o externa. Las representaciones externas se definen para la comunicación humano—máquina, y la interna, para la comunicación entre las máquinas [33].

Los grafos conceptuales para representar texto fueron introducidos por Sowa [33]. Los grafos conceptuales manejan dos tipos de nodos: conceptos y relaciones conceptuales. Los conceptos tienen un tipo (clase de concepto) y un referente (la instancia de este tipo de objeto). Las relaciones conceptuales señalan la manera en que los conceptos se relacionan [34]. Cada relación conceptual tiene uno o más (usualmente dos) arcos, cada uno de los cuales debe estar enlazado a un concepto [33].

Dado que la representación por medio de un grafo conceptual denota los términos que contribuyen a la semántica de la sentencia y que cada término se escoge de acuerdo a la posición dentro de la sentencia [35], los grafos conceptuales cuentan con una serie de características que hacen que sean muy ricos semánticamente y se utilicen no solo para el intercambio de información sino para la creación de bases de conocimiento y ontologías.

5.2 Formato de intercambio de conocimiento

El Formato de Intercambio de Conocimiento (*KIF*: *Knowledge Interchange Format*) se basa en caracteres que pueden ser combinados en lexemas; los lexemas a su vez pueden ser combinados en expresiones. La sintaxis del KIF [47] generalmente se presenta con una modificación de la notación de las formas BNF (*Backus-Naur forms*).

El alfabeto de KIF consiste de seis bloques de datos para referenciar las mayúsculas, las minúsculas, los dígitos, los caracteres alfa —caracteres específicos que se usan de la misma forma que las letras—, los caracteres especiales y otros caracteres como el espacio.

El proceso de convertir los caracteres a lexemas se llama análisis léxico. Al proceso entra una cadena de caracteres y se obtiene una cadena de lexemas. Este proceso es cíclico: en este proceso se leen las cadenas de caracteres hasta que se encuentra un carácter que no puede ser combinado con los caracteres previos y dentro del lexema actual. Cuando esto ocurre, el proceso se vuelve a empezar con el nuevo carácter y otro lexema.

KIF maneja cinco clases de lexemas: lexemas especiales, palabras, referencias al carácter, cadenas de caracteres y bloques de caracteres.

Los lexemas se forman de acuerdo a una serie de reglas. Se presentan tres tipos de expresiones: términos, sentencias y definiciones. Los términos son usados para denotar objetos, las sentencias para expresar hechos y las definiciones para definir constantes. Las definiciones y las sentencias se llaman formas; una base de conocimiento es un conjunto finito de formas.

La base de la semántica de KIF es la conceptualización del mundo en términos de objetos y relaciones entre los objetos. El universo del discurso es el conjunto de todos los objetos que hipotéticamente existen en el mundo. La noción de objeto es amplia. Los objetos pueden ser concretos o abstractos, primitivos o compuestos y pueden ser de ficción.

Además de permitir incluir listas, el lenguaje permite incluir sentencias matemáticas, de control, de relaciones y lógicas.

5.3 Infraestructura para la descripción de recursos

El leguaje RDF (*Resource Description Framework*) [48] ha sido definido para representar información sobre recursos en la Web. En particular, intenta representar metadatos sobre los recursos de la Web como el título, el autor, fechas, derechos y en general cualquier información relevante. Por otro lado, proporciona interoperabilidad entre las diferentes aplicaciones que intercambian información en la Web. Su desarrollo se ha basado no solo en las necesidades de la Web sino en los demás estándares que definen las diferentes comunidades, tales como los presentados a continuación.

Lenguaje de metadatos para publicar hipertexto en Internet. El lenguaje HTML (*HyperText Markup Language*) es estandarizado por el grupo W3C y es el más popular para escribir las páginas Web. Permite describir la estructura y el contenido en forma de texto, incluir imágenes, tablas, vínculos y muchos otros aspectos de presentación y diseño.

Plataforma para la selección de contenido en Internet. Esta especificación (*PICS*: *Plataform for Internet Content Selection*) [49] habilita los metadatos que pueden ser asociados con el contenido de Internet. PICS fue inicialmente diseñada para ayudar al control de contenido que acceden los menores de edad en la Web; sin embargo actualmente es ampliamente utilizada en los filtros. En general este estándar permite etiquetar el contenido propio o relacionado, creando así el principal parámetro de control.

Lenguaje de marcado generalizado. El lenguaje SGML (*Standard Generalized Markup Language*) se define como un sistema para la organización y etiquetado de documentos. Al igual que el HTML, fue normalizado por la Organización Internacional de Estándares (ISO) en 1986.

Lenguaje de marcado extensible. El Lenguaje de Marcado Extensible (*XML*: *Extensible Markup Language*) es un metalenguaje extensible de etiquetas desarrollado por el W3C. Es una simplificación y adaptación del SGML. Permite definir la gramática de lenguajes específicos, como HTML. En general XML no es un lenguaje en particular, sino una manera de definir lenguajes para diferentes necesidades. XML estandariza el intercambio de información estructurado entre las diferentes plataformas computacionales.

5.4 Lenguaje ontológico

El Lenguaje de Ontologías para la Web (*OWL*: *Web Ontology Language*) [50] es un lenguaje de marcado desarrollado por el grupo W3C para publicar y compartir ontologías en la Web. Fue desarrollado como una extensión del RDF y del lenguaje de marcado semántico para recursos en la Web DAML+OIL fusión de los lenguajes (*DARPA*[1] *Agent Markup Language)* DAML y (*Ontology Inference Layer or Ontology Interchange Language*) OIL [43].

1 Defense Advanced Research Projects Agency (DARPA)

Una ontología OWL es un grafo RDF que permite expresar relaciones complejas entre las diferentes clases de RDFs. Provee los recursos para determinar propiedades y elementos y para construir nuevas clases a partir de otra u otras.

5.5 Lógica común

Lógica Común (*CL*: *Common Logic*) es una estructura definida para una familia de lenguajes lógicos basados en lógica de primer orden. Define estándares para el intercambio de información basados en formas sintácticas llamadas dialectos. Un dialecto puede usar cualquier sintaxis que conforme una semántica abstracta CL. Todos los dialectos son equivalentes, es decir que pueden automáticamente ser traducidos entre ellos aunque algunos pueden ser más expresivos que otros, en cuyo caso se pueden traducir solo a menos expresivos. El estándar ISO 24707 para la Lógica Común especifica tres tipos de dialectos:

- Formato de Intercambio de Lógica Común (*CLIF*: *Common Logic Interchange Format*),
- Formato de Intercambio de Grafos Conceptuales (*CGIF*: *Conceptual Graph Interchange Format*),
- Notación basada en XML para la Lógica Común (*XCL*: *XML based notation for Common Logic*).

Son muchos los lenguajes que hereden de una sintaxis abstracta de la CL, entre los cuales destaca el que se presenta en [30].

Lenguajes naturales controlados. Los lenguajes naturales controladas son subconjuntos de los lenguajes naturales restringidos en la gramática y el vocabulario con el fin de reducir o eliminar la ambigüedad y la complejidad [31]. Los lenguajes controlados pueden ser desarrollados con dos objetivos: aquellos que mejoran la legibilidad para los lectores humanos y aquellos que permiten el análisis semántico automático confiable del texto. Dentro de estos lenguajes están el inglés, el chino y el español controlados [30].

Diagramas FLIPP (*Format for Logical Information Planning and Presentation***).** Se conocen como una representación lógica conceptual que no hace uso de texto ni símbolos. Cada diagrama consiste en un conjunto de bloques dependientes a nivel jerárquico [54]. El diagrama total represente un grafo acíclico. Cada sub-bloque puede contener información en lenguaje declarativo, natural o matemático.

Mapas de tópicos y mapas conceptuales. Los mapas conceptuales son artefactos para la organización y representación del conocimiento. Tienen su origen en teorías sobre psicología [51]. El objetivo de estos mapas es de representar relaciones entre conceptos en forma de proposiciones. Los conceptos están incluidos en cajas o círculos, mientras que las relaciones entre ellos se explicitan mediante líneas que unen las cajas respectivas. Las líneas, a su vez, tienen palabras asociadas que describen la naturaleza de la relación que liga los conceptos.

Lenguaje de modelamiento unificado[2]. Es un lenguaje de modelado de software (*UML*: *Unified Modeling Language*). Su estándar es definido por el grupo *Object Management*

2 Ver http://www.omg.org

Group (OMG). El estándar define el modelo estático y dinámico de todos los componentes que pueden hacer parte de un sistema de software incluidos los referentes a las ontologías para la representación, manejo e interoperabilidad y aplicaciones para la semántica de los negocios.

Otros lenguajes. El lenguaje de consulta estructurado (SQL: *Structured Query Language*[3]) es un lenguaje declarativo que permite recuperar información estructurada de las bases de datos relacionales. El lenguaje de restricción de objetos (OCL: *Object Constraint Language*[4]) definido por el grupo OMG, para describir las reglas que aplican al UML. Prolog[5] es un lenguaje de programación lógico e interpretativo. Datalog, una derivación del Prolog, es un lenguaje de consulta para bases de datos deductivas. Esquema RDF, una extensión semántica de RDF, es un lenguaje primitivo para la descripción de vocabulario ontológico.

6. Conclusiones

La búsqueda de nuevas alternativas para la representación y procesamiento del lenguaje natural que permitan no solo la inclusión de la semántica propia del lenguaje sino que faciliten operaciones que lleven a la inferencia del conocimiento, es un tema que aún está en espera de mostrar resultados.

Como se puede concluir de la revisión presentada, si bien existen muchas estructuras, a mayor cantidad de elementos semánticos que se pueden incluir, mayor es su dificultad para el procesamiento. Gran parte del camino a recorrer con estas estructuras aún esta por explorar, no solo en lo que se refiere a la representación del lenguaje natural, sino igualmente, a la creación de ontologías y bases de conocimiento.

Agradecimientos. El trabajo fue realizado parcialmente con el soporte que el segundo autor recibió del Gobierno de México (CONACYT 50206-H, SIP-IPN 20113295, programa CONACYT de estancias Sabáticas 2010-2011) y de Universidad Waseda, Japón.

7. Referencias

[1] Salton, G. y Lesk, M. E. (1965). The SMART automatic document retrieval systems and illustration. *Commun. ACM.*

[2] Farkas, J. (1966). Improving the classification accuracy of automatic text processing systems using context vectors and back-propagation algorithms. *Canadian Conference on Electrical and Computer Engineering.*

[3] Henderson, J., Merlo, P., Petroff, I. y Schneider, G. (2002). Using NLP to efficiently visualize text collections with SOMs. *Proceedings: 13th International Workshop on Database and Expert Systems Applications.*

[4] Kimura, M., Saito, K., Ueda, N. (2005). Multinomial PCA for extracting major latent topics from document streams. *Neural Networks 2005, IJCNN '05. Proceedings. 2005 IEEE International Joint Conference.*

[5] Maron, M.E. y Kuhns, J.L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM.*

[6] Badia, A. y Kantardzic, M. (2005). Graph building as a mining activity: finding links in the small. *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery.* 17-24. ACM.

3 En las especificaciones ISO/IEC 9075-1:2008; ISO/IEC 9075-2:2008; ISO/IEC 9075-11:2008, se incluyen los mínimos requerimientos del lenguaje.
4 Ver http://www.omg.org
5 ISO/IEC 13211-1 ISO-Prolog

[7] Rijsbergen C.J. van, Robertson S.E. y Porter M.F. (1980). New models in probabilistic information retrieval. *London: British Library. (British Library Research and Development Report,* No. 5587.

[8] Klabbankoh B. y Pinngern Q. (2000). Applied Genetic Algorithms in Information Retrieval. *Faculty of Information Technology, King Mongkut´s Institute of Techology Ladkrabang*.

[9] Varlamis I., Vazirgiannis M., Halkidi M., Nguyen B. (2004). Thesus, A Closer View on Web Content Management Enhanced with Link Semantics. *IEEE Transactions on Knowledge and Data Engineering,* Vol. 16. No. 6. 685-700.

[10] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K. y Harshman R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science,* Vol. 41. No. 6. 391-407.

[11] Hensman S. (2004). Automatic Construction of Conceptual Graphs from Texts using Computational Linguistics Techniques. *Department of Computer Science, University College Dublin. Belfield, Dublin 4. Proceedings of Student Research Workshop at HLT-NAACL*.

[12] Rijsbergen Van C.J. (1979). Information Retrieval. *Department of Computing Science, University of Glasgow Second edition.*

[13] Schenker A., Bunke Horst, M. L. A. K. (2005). Graph-theoretic techniques for Web content mining. *World Scientific Publishing.*

[14] Schenker A., Bunke H., M. L. y Kandel, A. (2004). A Graph-Based framework for Web document mining. *Lecture Notes in Computer Science Publisher Springer Berlin Heidelberg,* Vol. 3163. 401-412.

[15] Barski C. (2009). The enigmatic art of knowledge representation. Consultado: www.lisperati.com//ex.html. (5 de marzo, 2009).

[16] Losada, D. y Barreiro A. *(*2001). Rating the impact of logical representations on retrieval performance. *Database and Expert Systems Applications Proceedings. 12th International Workshop,* 247-253.

[17] Chang C. y Chen A. (1998). Supporting conceptual and neighborhood queries on WWW. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions,* Vol. 28. No. 2. 300-308.

[18] Peltonen, J., Sinkkonen J. y Kaski, S. (2002). Discriminative clustering of text documents. *9th International Conference Neural Information Processing, 2002. ICONIP '02,* Vol. 4. 1956-1960.

[19] Maisonnasse L., Gaussier E., J. C. (207). Multiplying Concept Sources for Graph Modeling. *LIG contribution to the CLEF 2007 medical retrieval task (ImageCLEFmed).*

[20] Gao J., Nie J.-Y, Wu G., Cao G. (1999). Dependence Language Model for Information Retrieval. *Microsoft Research, Asia, Brooks Cole Publishing Co., Pacific Grove*.

[21] Williams R. (2007). A Computational Effective Document Semantic Representation. *DEST'07. Digital EcoSystems and Technologies Conference, IEEE-IES*.

[22] Manning C., Schütze H. (1999). Foundations of Statistical Natural Language Processing. *MIT Press. Cambridge, MA: May*.

[23] Mineau, G. W., Stumme, G. y Wille, R. (1999). Conceptual Structures Represented by Conceptual Graphs and Formal Concept Analysis. *International Conference on Conceptual Structures*.

[24] Delugach H. S. Towards. (2008). Conceptual Structures Interoperability Using Common Logic Computer. *Science Department Univ. of Alabama in Huntsville. Third Conceptual Structures Tool Interoperability Workshop*.

[25] Helbig H. (2006). Knowledge Representation and the Semantics of Natural Language. *Lecture Notes in Computer Science. Springer.*

[26] Sowa, J. F. (2008). Conceptual Graphs. *Handbook of Knowledge Representation*.

[27] Gelbukh A., Sidorov G. (2006). Procesamiento automático del español con enfoque en recursos léxicos grandes. *Instituto Politécnico Nacional, México*.

[28] Hernández Cruz, M. (2007). Generador de los grafos conceptuales a partir del texto en español. *Tesis de Maestría. Instituto Politécnico Nacional. Centro de Investigación en computación*.

[29] Last M., Maimon O. (2004). A compact and Accurate Model for Classification. *IEEE Transactions on Knowledge and Data Engineering.,* Vol. 16. No. 2. 203-215.

[30] Sowa, J. F. (2008). Common Logic, A Framework for a Family of Logic-Based Languages.

[31] Barceló, G., Cendejas, E., Bolshakov, I. y Sidorov G. (2009). Ambigüedad en nombres hispanos. *Revista Signos. Estudios de Lingüística 42 (70).* 153-169.

[32] Shin S-J. (1994). The Logical Status of Diagrams. *Cambridge University Press.*

[33] Committee on Information Interchange and Interpretation. Sowa, J. F. (2008). *Conceptual Graph Standard.* Consultado: www.jfsowa.com/cg/cgstandw.htm. (12 de noviembre, 2008).

[34] Montes-y-Gómez M. (2001). Minería de texto: Un nuevo reto computacional. *3er Taller Internacional de Minería de Datos MINDAT-2001, Universidad Panamericana, Ciudad de México*.

[35] Shehata, S., Karray, F. y Kamel, M. (2006). Enhancing Text Retrieval Performance using Conceptual Ontological Graph. *Data Mining Workshops,* 39-44*, Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06).*

[36] Tesnière, A. L. (1959). Elements de syntax e structurale. *Klincksieck Paris.*

[37] Castro-Sánchez, N. A., y Sidorov, G. (2010). Analysis of Definitions of Verbs in an Explanatory Dictionary for Automatic Extraction of Actants based on Detection of Patterns. *Lecture Notes in Computer Science,* No. 6177. 233-239.

[38] Abdulrub S., Polovina S. y Hill, R. (2008). Implementing Interoperability through an Ontology Importer for Amine. Conceptual Structures Tools and the Web - Third Conceptual Structures Tool Interoperability Workshop.

[39] Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal,* Vol. 27, 379–423.

[40] Sleator, D. D. y Temperley, D. (1993). Parsing English with a link grammar. Third International Workshop on Parsing Technologies.

[41] Roberts R, Goldstein I,. (1977). *T*he FRL manual. *Memo 409, Massachusetts Institute of Technology, Artificial Intelligence Laboratory*.

[42] Noy N., Grosso W, Musen A. (2000). Knowledge acquisition Interfaces for Domain Experts: An Empirical Evaluation of Protege-2000. *Twelfth International Conference on Software Engineering and Knowledge Engineering (SEKE 2000), Chicago, IL.*

[43] McGuinness D. L., Fikes R., Hendler J., Stein L. A., (2002). DAML+OIL: An Ontology Language for the Semantic Web. *IEEE Intelligent Systems,* Vol. 17, No. 5. 72-80, Sep./Oct.

[44] Kifer M, Lausen G, Wu J. (1995). Logical Foundations of Object-Oriented and Frame-Based Languages. *Journal of ACM, May.*

[45] Brachman R., Schmolze J. (1985). An overview of the KL-ONE Knowledge Representation System. *Cognitive Science*, Vol. 9, No. 2, 171-216.

[46] Baader F., Nutt W. (2002). Basic Description Logics. *The Description Logic Handbook, Cambridge University Press.* 47-100.

[47] Ginsberg M. (1991). Knowledge Interchange Format: The KIF of Death. *Journal AI Magazine,* Vol. 12. 57-63.

[48] Lassila O., Swick R. y World Wide and Web Consortium. (1999). Resource Description Framework (RDF) Model and Syntax Specification. *Consortium, Cambridge (MA). W3C Recommendation.*

[49] Krauskopf T., Miller. J, Resnick P., Treese W. (1996). PICS Label Distribution Label Syntax and Communication Protocols. *Version 1.1, W3C Recommendation.*

[50] Sirin E., Hendler J., Parsia B. (2002). Semi-automatic Composition of Web Services using Semantic Descriptions. *Web Services: Modeling, Architecture and Infrastructure workshop in ICEIS 2003.* 17-24.

[51] Ausube l, D., Novak, J. Hanesian, H. (1978). Psicología Educacional: Una visión cognitiva. *Halt, Reinhart and Winston, New York.*

[52] Bally C, Sechehaye C. (1945). Curso de Linguistica General. *Editorial losada, Buenos Aires.* 31-32.

[53] Chomsky, N. (1957). Syntactic structures. *La Haya, Mouton.*

[54] Croitoru M., Jäschke R. (2008). Conceptual Structures Tools and the Web. *Third Conceptual Structures Tool Interoperability Workshop.*

Sonia Ordoñez Salinas

Docente Universidad Distrital – Facultad de ingeniería. Estadística de la Universidad Nacional. Ingeniera de Sistemas de la Universidad Distrital. Especialista Teleinformática Universidad Distrital. Magíster en Ingeniería de Sistemas Universidad Nacional. Doctor Ing. Sistemas y Computación, Universidad Nacional de Colombia. Grupo de Investigación Gesdatos U.D.

Alexander Gelbukh

Profesor-Investigador y Jefe del Laboratorio de Lenguaje Natural y Procesamiento de Texto del Centro de Investigación en Computación del Instituto Politécnico Nacional, México. Doctor en la ciencia de la computación por el Instituto de la Información Científica y Técnica de toda Rusia (VINITI). Maestro en Ciencias en matemáticas por la Universidad Nacional "Lomonósov" de Moscú (MGU), Rusia. Miembro de la Academia Mexicana de Ciencias, Investigador Nacional de México nivel II.