Universidad Complutense de Madrid

Facultad de Filología

Grado en Lingüística y Lenguas Aplicadas

Trabajo de Fin de Grado

UNA REVISIÓN DE LOS MODELOS COMPUTACIONALES DE REPRESENTACIÓN DEL SIGNIFICADO INTRÍNSECO DEL TEXTO PARA SU PROCESAMIENTO AUTOMÁTICO

Elisa Caballero Redondo

Tutora: Ana Fernández-Pampillón Cesteros

Junio de 2017

Una revisión de los modelos computacionales de representación del significado intrínseco del texto para su procesamiento automático.

Resumen

El problema general que plantea este trabajo es cómo representar el significado intrínseco de los textos para poder ser procesados automáticamente. El significado intrínseco se refiere a la construcción de estructuras de conocimiento que representen los conceptos que contiene el texto sin referirse a conocimiento externo o conocimiento "del mundo" como es, por ejemplo, el contenido en ontologías o bases de conocimiento. Para encontrar una solución, se ha realizado una revisión general del estado de la cuestión en Lingüística Computacional. Dada la amplitud de soluciones encontradas que están específicamente orientadas a un dominio de conocimiento particular o a un tipo de procesamiento del texto específico, se han seleccionado únicamente los mecanismos de representación que, con mayor frecuencia, se utilizan en los sistemas de Procesamiento del Lenguaje Natural actuales con el fin de ofrecer una síntesis lo más general e independiente del contexto posible.

1. Introducción

El problema que aborda este trabajo es cómo representar el significado intrínseco de los textos para poder ser procesados automáticamente. Por representación del significado intrínseco nos referimos a la construcción de estructuras de conocimiento que representen los conceptos que contiene el texto sin utilizar conocimiento externo como el contenido en ontologías o bases de conocimiento. Esto último implica que sería necesario construir ese conocimiento externo sobre el mundo antes de obtener las representaciones textuales, lo cual lleva mucho más trabajo e implica tener muchos más recursos (Allen, 1995). Por este motivo, y como primera aproximación a la representación computacional del texto, no lo vamos a considerar en este trabajo.

La representación del significado textual es uno de los problemas básicos de la Lingüística Computacional, del Procesamiento del Lenguaje Natural (PLN) y de la Lingüística Textual. En este sentido, es una cuestión que necesita de una perspectiva multidisciplinar que recoja las propuestas formuladas en Lingüística, Inteligencia Artificial, Psicología Cognitiva y las Neurociencias entre otras.

La elección del texto como objeto de este trabajo no es arbitraria. Se puede afirmar que el texto es el formato estándar de la información en los entornos digitales, y, en todo caso, el formato más sencillo de procesar. En el año 2007 se calculaba que, si toda la información almacenada en Internet se comprimía óptimamente, se obtendría un total de unos 300

exabytes (1018 bytes), de los cuales, aproximadamente un 20% correspondía a texto digital (Hilbert, 2014). Dicho porcentaje ha ido aumentando de forma progresiva desde 1986 (cuando era tan solo un 0,2% del total), haciéndose cada vez más importante con respecto a otros formatos como audio y vídeo. Esto significa que el procesamiento de textos (el tratamiento automático de los textos para la obtención de información y conocimiento) tiene, y tendrá previsiblemente, un gran interés científico y económico.

Así, la motivación de este trabajo surge de la necesidad de obtener información a partir de cantidades ingentes de datos textuales con la ayuda de las máquinas. Esta tarea, que es inabordable para la personas, podría lograrse si las máquinas fuesen capaces de representar y procesar el texto como lo haría un ser humano. Esta posibilidad, sin embargo, sigue siendo una cuestión, que a día de hoy, no está totalmente resuelta a nivel global aunque sí de forma parcial considerando la representación de sólo algunos aspectos del texto (Feldman & Sanger, 2007).

Comprender correctamente un texto es complejo, incluso para nosotros, debido, fundamentalmente, a que es necesario utilizar gran cantidad de conocimiento del mundo (conocimiento común) y conocimiento lingüístico (Rich & Knight, 1998). En este sentido, uno de los objetivos de la Lingüística Computacional es diseñar y probar la eficacia de modelos de representación semántica de carácter simbólico, como las formas lógicas (Allen, 1995), de carácter estocástico, como la Bolsa de Palabras (Serrano Moreno, 2007), o de carácter biológico, como las redes neuronales (Fausett, 1994). Estos modelos constituyen la base para construir aplicaciones para extraer automáticamente información o datos del lenguaje natural como, por ejemplo, las aplicaciones de Extracción de Datos, Análisis de Opinión o Clasificación de Textos (Feldman & Sanger, 2007), o bien para lograr una comunicación satisfactoria entre humanos o entre humanos y máquinas como en el caso de los traductores automáticos (Feldman & Sanger, 2007) o, los actualmente muy demandados, *chatbots* (Yao, 2017).

Para llevar a cabo la representación textual existen diferentes propuestas que son más o menos adecuadas según la tarea a resolver. Por ello, el problema no es sólo conocerlas sino también saber seleccionar la más adecuada a un determinado problema de procesamiento. En este sentido, saber seleccionar el modelo de representación adecuado es importante porque de la correcta selección depende que el modelo sea eficaz, es decir, que la respuesta que proporcione sea válida, y eficiente, es decir, que necesite el mínimo tiempo de procesamiento y los mínimos recursos informáticos posibles.

Este Trabajo de Fin de Grado pretende contribuir a facilitar la selección del modelo de representación textual más adecuado para una determinada tarea de procesamiento y, para ello, se ha realizado una revisión de los modelos más paradigmáticos de la representación del significado intrínseco de un texto, es decir, la interpretación de un texto sin considerar la existencia o la necesidad de referirse a ningún tipo de conocimiento previo o externo al texto.

El trabajo se ha estructurado en cinco secciones. La sección actual, la primera, ha introducido la cuestión y la motivación del trabajo. La segunda sección describe la metodología de trabajo. La tercera sección presenta una síntesis sobre las cuestiones básicas a tener en cuenta para la representación del texto: su definición, caracterización, coordinación y la composicionalidad. La cuarta sección sintetiza la selección de los seis modelos computacionales de representación intrínseca textual. Finalmente, en la quinta sección se presentan las conclusiones.

2. Metodología del trabajo

El primer paso de la metodología seguida para la revisión del estado de la cuestión del problema planteado fue la selección y búsqueda del material bibliográfico. La primera búsqueda se hizo utilizando los libros recomendados en la bibliografía de las asignaturas de Lingüística Textual, Lingüística Computacional y Procesamiento del Lenguaje Natural del Grado de Lingüística y Lenguas Aplicadas que se referían a Representación Textual o Minería de Textos. El segundo paso fue la realización de un análisis centrado en qué modelos podían servir para resolver el problema planteado. El tercer paso fue la realización de una segunda búsqueda más precisa y orientada al problema para ampliar información. Esta información se analizó y se seleccionó lo que era relevante para este trabajo. En el cuarto paso, toda esa información se sintetizó y se estructuró de tal forma que coincidiera con el punto de vista que se quería mostrar. Finalmente, el quinto paso fue la escritura del trabajo. La bibliografía, recogida al final del trabajo, consta de 24 obras incluyendo tanto libros como artículos.

3. Representación del significado intrínseco del texto

3.1. Caracterización del texto

El concepto de texto no es un concepto fácil de definir. Los expertos en Lingüística Textual no terminan de ponerse de acuerdo a la hora de dar una definición exacta de lo que es texto.

En este trabajo se considerará la definición que aporta Enrique Bernárdez (Bernárdez, 1983:85).

> “”Texto’ es la unidad lingüística comunicativa fundamental, producto de la actividad verbal humana, que posee siempre carácter social; está caracterizado por su cierre lingüístico y comunicativo, así como por su coherencia profunda y superficial, debida a la intención (comunicativa) del hablante de crear un texto íntegro, y a su estructuración mediante dos conjuntos de reglas: las propias del nivel textual y las del sistema de la lengua.”

Así como la definición propuesta por Lázaro Carreter (Carreter, 1971:391) que añade, respecto a la anterior, qué fragmentos del lenguaje natural pueden considerarse un texto:

> “Son textos, por lo tanto, un fragmento de una conversación, una conversación entera, un verso, una novela, la lengua en su totalidad, etc.”

Texto es un concepto que abarca, por lo tanto, algo tan pequeño como puede ser una frase hasta algo tan grande como un libro o una colección de libros. Al incluir material de extensiones tan diversas, hay que tener en cuenta que para analizar, comprender y representar el texto es necesario contar con diversos niveles de conocimiento lingüístico (Allen, 1995). El nivel morfológico, el nivel sintáctico, el nivel semántico y el nivel discursivo son los que intervienen, fundamentalmente, en estas tareas. Es decir, hay que valorar el significado de las palabras, cómo se combinan para formar oraciones y la correcta interpretación del resultado. El nivel pragmático y el nivel de conocimiento del mundo no se tendrán en cuenta pues ambos niveles vienen determinados por conocimiento almacenado en sistemas externos al sistema de procesamiento de lenguaje natural como son las bases de conocimiento (eg. ontologías) y, como se ha apuntado anteriormente, en este trabajo no se van a considerar.

En el texto, desde el punto de vista lingüístico (Cuenca, 2010), se pueden distinguir dos partes complementarias. Por un lado, la parte interpretativa del texto que viene definida por la adecuación y la cohesión; por otro lado, la parte formal que viene determinada por la coherencia del texto: de qué habla, qué información aporta y cómo está organizada. En este trabajo el interés se centra en conocer cómo se obtiene el significado intrínseco del texto lo que se corresponde con la parte de la coherencia.

En este sentido, conviene tener en cuenta que la coherencia posee varios elementos constitutivos:

- Tema del texto: es el contenido básico del texto.

- Patrón estructural: según los modelos convencionales del tipo de texto en cuestión. Por ejemplo, si el texto es narrativo, tendrá una estructura en forma de planteamiento, nudo y desenlace.
- Selección de información: se busca focalizar la atención en ciertos aspectos temáticos en función de la relevancia respecto al tema o la intención comunicativa.
- Organización de la información: es la ordenación jerárquica de la información (conocida y nueva) del texto.

Así, por ejemplo, estos elementos de coherencia se pueden identificar en el siguiente texto ilustrativo (Fuente: Europa Press, 10 de agosto de 2012) con título "Valentino Rossi vuelve a Yamaha para las dos próximas temporadas":

> "El piloto italiano Valentino Rossi abandonará la próxima temporada el equipo Ducati, donde ha encadenado decepciones los dos últimos años, para volver a Yamaha, con la que ha firmado un contrato por dos temporadas y donde volverá a compartir equipo con el vigente líder del Mundial de MotoGP, Jorge Lorenzo.
>
> Ducati anunció este viernes que Rossi abandonaría el quipo y Yamaha confirmó acto seguido la vuelta del italiano para las temporadas 2013 y 2014. 'Il Dottore' intentará reverdecer viejos laureles en la marca del diapasón, con la que ganó cuatro mundiales y 46 carreras de la máxima categoría en una fructífera etapa que se prolongó siete años entre 2004 y 2010.
>
> Por su parte, Ducati ha subrayado que ya está trabajando en el nuevo proyecto, después de "haber renovado recientemente" al estadounidense Nicky Hayden."

Este fragmento se compone de cuatro oraciones que están relacionadas entre sí de forma coherente ya que se refieren a un tema concreto: "la vuelta del piloto Valentino Rossi al equipo Yamaha."

Al ser un texto dentro del ámbito periodístico (una noticia), su estructura está marcada por el patrón de este tipo de textos. La información se selecciona de manera que lo que aporte sea contenido relevante al tema que se trata y dicha información se organiza de manera lógica. Primero se sitúa el tema ("el piloto abandonará el equipo Ducati"), es decir, se da la noticia principal, y luego se explican las causas de la vuelta al equipo Yamaha, exponiendo que el piloto ya estuvo en este equipo y ahora vuelve a él. Esto es, se aporta información nueva e información ya conocida.

Como se puede observar, para construir el texto se parte de cuatro oraciones independientes que al unirse y cohesionarse forman un texto. Pero, ¿cómo se transcienden los límites de la oración para llegar a ese texto? Es decir, ¿qué es lo que da a la cohesión de unas oraciones el sentido de unidad textual? La respuesta es la coordinación (Bernárdez, 1982), que se presenta en la siguiente subsección 3.2, relacionada directamente con el Principio de Composicionalidad, que se describe en la subsección posterior 3.3.

3.2. La coordinación

El concepto de coordinación recibe diversas denominaciones según las distintas escuelas. En el presente trabajo se adoptará la denominación de Marco de Integración Global propuesta por el lingüista Ewald Lang (Bernárdez, 1982:144).

> "Las operaciones realizadas mediante el significado operativo de las conjunciones son operaciones sobre los significados de los conjuntos, con ayuda de las cuales se ponen en mutua relación los significados de las oraciones, por medio de la reflexión, con el resultado de que, a partir de los significados de las oraciones […], se constituye una unidad distinta a los significados de los conjuntos, que llamo Marco de Integración Global"

Para entender la diferencia entre una sucesión de oraciones sin coherencia que no adquiere la categoría de texto y una sucesión de oraciones que sí pueden ser consideradas texto, se ilustrará con los siguientes ejemplos.

"Vivo en un piso en el centro de Madrid. El cielo es azul. Mi gato se llama Guantes"

"Vivo en un piso en el centro de Madrid. En Madrid los pisos del centro suelen ser pequeños pero acogedores. Es uno de los mejores en los que he vivido."

El primer ejemplo no se podría llegar a considerar texto pues no existe un Marco de Integración Global que las una, es decir, no hay ninguna conexión posible con coherencia porque no existe relación semántica entre ellas. Sin embargo, el segundo ejemplo sí que podría constituir un texto ya que el Marco de Integración Global lo dota de unidad, de significado conjunto y de coherencia, pues se pueden establecer relaciones semánticas entre las frases que lo constituyen. Una sucesión de oraciones nunca podrá considerarse texto si carece de coherencia. Es necesario que exista un núcleo, asunto o plan global del texto, es decir, es necesario que exista un Marco de Integración Global. Por lo tanto, la coordinación

es un requisito necesario para que una sucesión de oraciones pueda llegar a ser un texto coherente.

3.3. La composicionalidad

El concepto de coordinación enlaza directamente con el Principio de Composicionalidad. Este Principio fue expuesto por Frege (Frege, 1948). Para él, el significado de una oración se construye a partir del significado de las partes que la constituyen. Para que la oración tenga sentido, también deben tenerlo dichas partes.

Si se une este Principio con la idea de la coordinación, el significado global del texto dotado de coherencia se construirá a partir del significado de las partes que lo constituyen, es decir, sus oraciones. Esta noción es fundamental en este trabajo. Tres de los seis modelos computacionales para la representación intrínseca del significado del texto que se presentarán (representaciones basadas en Lógica, Redes Semánticas y Marcos), son modelos capaces de representar el significado de oraciones. Si dichas oraciones son parte de un texto corto, estos modelos necesitarán de este Principio para unir las representaciones de estas oraciones para lograr la representación del texto. Es decir, se necesita sumar las representaciones de los componentes del texto para lograr la representación del texto en sí.

En la próxima sección se presentarán los modelos para la representación intrínseca del significado del texto y se profundizará más en la aplicación del Principio de Composicionalidad. No obstante, como se verá a continuación, este Principio genera problemas al aplicarlo en la representación de textos de tamaño medio y grande.

4. Modelos computacionales de representación del significado intrínseco del texto

En este apartado se presentarán los seis modelos computacionales seleccionados como paradigmáticos de representación del significado intrínseco del texto. Han sido seleccionados por ser modelos generales (independientes del dominio de conocimiento) y ampliamente utilizados por los sistemas de procesamiento textual a lo largo de la historia de la Lingüística Computacional, lo que implica cierta independencia de la aplicación de procesamiento. Los primeros tres modelos, representaciones del significado basadas en Lógica, en Redes Semánticas y en Marcos, se aplican para representar textos cortos: a nivel de frase y a nivel discursivo local. Los tres modelos siguientes, representaciones del significado en Árbol y

Segmentos, Bolsa de Palabras y Análisis Semántico Latente, se utilizarán para representar textos de tamaño medio y grande, a nivel discursivo global.

Al hablar del nivel de frase se habla de la interpretación de una frase aislada, sin tener en cuenta ni la frase anterior ni la posterior (Allen, 1995). El nivel discursivo local, por su parte, tiene en cuenta la frase a interpretar, la frase anterior y la posterior. El nivel discursivo global va un paso más allá y tiene en cuenta todas las frases de un texto, agrupándolas en unidades, llamadas segmentos, para realizar la interpretación.

4.1. Representaciones del significado basadas en la Lógica

4.1.1. Representaciones a nivel de frase

La Lógica formal se desarrolló como una notación formal para capturar las propiedades esenciales del lenguaje natural y es uno de los modelos más utilizados para representar la interpretación final de las oraciones en Lingüística Computacional. Al ser un lenguaje formal, los ordenadores son capaces de almacenarlo y manipularlo (Allen, 1995). Esta sección se centrará en los dos tipos de lógica más básicos: la Lógica proposicional y la Lógica de predicados de primer orden.

La Lógica proposicional es un lenguaje que consta de tres tipos de elementos en el vocabulario: las variables, las constantes y las conectivas. Las variables son los símbolos que sustituyen cualquier proposición y, como su nombre indica, pueden cambiar dependiendo de la expresión. Se utilizan las letras *p, q, r, s, t, etc.* como variables de cada proposición. Por lo tanto, la variable *p* sustituiría a la proposición "*ahora llueve"* y puede tener tanto el valor de verdad como el valor de falsedad. Las constantes son los elementos cuyo significado no varía con cada interpretación y representan las relaciones existentes entre proposiciones. Las conectivas son constantes lógicas y su función es conectar proposiciones. Los más frecuentes son: negación ("*no llueve ahora"* se simbolizaría como $\neg p$), conjunción ("*hoy llueve y nieva"* sería $p \wedge q$), disyunción ("*o peras o manzanas"* sería $p \vee q$), condicional ("*si llueve, la tierra se moja"* se representaría como $p \rightarrow q$) y bicondicional ("*es de noche si y solo si se ha puesto el sol"* se simbolizaría como $p \leftrightarrow q$).

Sin embargo, la Lógica proposicional encuentra limitaciones a la hora de expresar el significado que se necesita recoger de una frase. Por ejemplo, no es posible representar frases como *"algunas manzanas son rojas"*. Esta afirmación no se refiere específicamente a ningún conjunto de manzanas, solo indica que existe un conjunto de manzanas que son rojas. La

Lógica de predicados de primer orden surge para solventar estos problemas al introducir operadores. Los operadores son elementos que combinan relaciones. El operador $\forall$ representa "para todo" y es la generalización de la conectiva $\wedge$ mientras que el operador $\exists$ representa "existe al menos uno" y es la generalización de la conectiva $\vee$. Estas conectivas son equivalentes a las funciones lógicas AND y OR respectivamente pues una indica conjunción y otra, disyunción. Así, la proposición "*algunas manzanas son rojas*" se representaría como $\exists x.\ M(x)$. Además, también se incluyen las conectivas de la Lógica proposicional. *"Nadie es sabio y además prudente"* se representaría como $\neg\exists x\ (Sx \wedge Px)$ y *"todos los sabios son prudentes"* como $\forall x\ (Sx \rightarrow Px)$.

También se pueden expresar razonamientos lógicos del tipo "todos los hombres son mortales. Sócrates es un hombre. Por lo tanto, Sócrates es mortal" que se representaría como:

$\forall x\ (Hx \rightarrow Mx)$ *("todos los hombres son mortales")* siendo:

$H(s)$ *("Sócrates es un hombre")*

$M(s)$ *("Sócrates es mortal")*

Este tipo de representación lógica es la base del lenguaje de programación Prolog ampliamente utilizado en el procesamiento del lenguaje natural (Clocksin & Mellish, 2003). En Prolog, un razonamiento de lógica de primer orden como $\forall x \forall y \forall z\ (progenitor(x, z) \wedge ancestro(z, y) \rightarrow ancestro(x, y))$ correspondería con: *ancestro(X,Y) :- progenitor(X,Z),ancestro(Z,Y).*

La Lógica proposicional y la Lógica de predicados tienen dos limitaciones importantes como sistemas de representación del lenguaje natural: la limitación del concepto de verdad (las expresiones o son falsas o son verdaderas, no admiten grados) y la limitación para reflejar la imprecisión del lenguaje (no existen grados de pertenencia de variables en conjuntos de entidades). Para solventar estas limitaciones, existen ampliaciones entre las que están la Lógica modal y la Lógica difusa.

La Lógica modal (Garson, 2013) introduce grados de verdad o falsedad mediante calificadores de posibilidad, imposibilidad, necesidad o contingencia pues las lógicas anteriores o bien son verdaderas o bien falsas. Por su parte, la Lógica difusa (Cintula, Fermüller, & Noguera, 2016) trata a las entidades en función de pertenencia a un conjunto impreciso. Por ejemplo en la frase *"Las anchoas son inteligentes"* (Serrano Moreno, 2007), el término inteligente es impreciso. La pertenencia de una anchoa al conjunto "inteligente" vendría determinada, por ejemplo, por su cociente intelectual. Así, si posee un cociente

intelectual de 210, tendría un grado de pertenencia de 0,9 al conjunto "inteligente", siendo 1 la pertenencia total. Si tuviera un cociente intelectual de 80, tendría un grado de pertenencia de 0,1 al mismo conjunto.

En definitiva, las distintas lógicas ofrecen mecanismos más o menos expresivos para representar formalmente y computacionalmente ciertos aspectos de la semántica de las frases, lo que permite el tratamiento automático de dichas representaciones por parte de sistemas inteligentes encargados de manipular estas representaciones como son, por ejemplo, los sistemas de gestión de bases de conocimiento o los traductores automáticos. No es posible, sin embargo, representar de forma completa toda la información que podría extraer un ser humano como el tiempo, modo, aspecto que se infiere de una frase en lenguaje natural.

4.1.2. Representaciones a nivel discursivo local

Se habla de la interpretación del nivel discursivo local a la interpretación que incluye tanto la información sintáctica y semántica de las frases inmediatas a la frase a analizar, es decir, la frase anterior y la frase posterior con el fin de interpretarla de forma correcta y completa.

Para lograr una representación a nivel discursivo local mediante la Lógica se usará el Principio de Composicionalidad anteriormente presentado.

En lo que respecta a la Lógica proposicional, aplicando el Principio de Composicionalidad, el primer paso para lograr la representación será dividir el texto del ejemplo en las oraciones que lo constituyen.

"Hacía frío en el interior de Moria y la oscuridad apenas era atravesada tímidamente por la luz de la antorcha. En un par de ocasiones se vio Gandalf en la necesidad de lanzar con el cayado, pero el brillo de los amenazadores ojos de criaturas extrañas y lejanas le hizo ser más prudente a cada minuto que pasaba entre las paredes de piedra."

Una vez realizado este paso, se procederá a representar cada frase.

"Hace frío en Moria" = p

"La antorcha alumbra poco" = q

"Gandalf alumbra con el cayado" = r

"El brillo de ojos de criaturas diversas le hace ser más producente" = s

Finalmente, se unirán y el resultado será la representación final a nivel discursivo local: *(p ∧ q) ∧ (r ∧ s)*

Como se puede observar, al ser un texto de tamaño pequeño, este modelo resulta efectivo pues es fácil seguir y realizar el razonamiento. Sin embargo, al intentar representar un texto de más tamaño, el resultado de la representación podría ser tan extenso que no resultaría útil, pues no se entendería.

Por su parte, la Lógica de predicados de primer orden sigue el mismo procedimiento. Al ser una lógica más expresiva que la Lógica proposicional, se ilustrará con un ejemplo de un texto más corto para demostrar que el resultado de la representación es menos manejable, pues es más extenso.

"Todos los amigos de Luis son amigos de Antonio. Sin embargo, los amigos de Pepe, que no son enfermeros, no lo son."

El siguiente paso, de nuevo, será representar cada oración por separado.

"Todos los amigos de Luis son amigos de Antonio" = $\forall x$ *[Amigo(x,Luis)* → *Amigo(x,Antonio)]*

"Los amigos de Pepe no son enfermeros" = $\forall x$ *[Amigo(x,Pepe)* ∧ ¬*Enfermero(x)*

"No lo son (amigos de Antonio") = ¬*Amigo(x,Antonio)]*

En último lugar, las representaciones de cada oración se unirán y dará como resultado la representación final a nivel discursivo local: $\forall x$ *[Amigo(x,Luis)* → *Amigo(x,Antonio)]* ∧ $\forall x$ *[Amigo(x,Pepe)* ∧ ¬*Enfermero(x)* → ¬*Amigo(x,Antonio)]*

En resumen, aplicar la Lógica a la representación a nivel de frase resulta útil pues permite el tratamiento automático de dichas representaciones. Sigue siendo útil para realizar representaciones a nivel discursivo local pero, a medida que se aumenta la información a representar, las representaciones finales son cada vez más extensas. Por lo tanto, si lo que se quiere es representar tanto oraciones como textos cortos, las representaciones son manejables pero si lo que se pretende es representar un texto de tamaño medio o grande, este modelo no resulta adecuado pues sería demasiado costoso descomponerlo todo entero en sus oraciones, realizar la representación de cada una y finalmente, unirlas. El resultado sería una representación final nada manejable al ser muy extensa.

4.2. Redes Semánticas

Una red semántica es un conjunto de nodos unidos por arcos que denotan asociaciones que pueden ser de diferentes tipos, dependiendo del dominio de conocimiento de los textos que se quieren representar. Existen tres tipos de redes asociativas, sin embargo, en este trabajo solo se va a utilizar el formalismo de Redes Semánticas para ilustrar. Los otros dos tipos de redes son las redes de clasificación y redes causales (Serrano Moreno, 2007).

En las Redes Semánticas (Collins & Quillian, 1969), el significado de un concepto depende de cómo se encuentre conectado con otros conceptos. El formalismo de representación del conocimiento en este modelo está basado en el formalismo matemático de los grafos. Un grafo es una estructura de organización de la información matemática formada por nodos conectados por arcos. La información se representa como un conjunto de nodos, utilizados para denotar conceptos o individuos o propiedades, conectados unos con otros mediante un conjunto de arcos etiquetados que representan las relaciones entre los nodos (eg. generalización/especialización). El mecanismo de inferencia básico en las Redes Semánticas es la herencia de propiedades a través de las relaciones de generalización/especialización.

4.2.1. Representaciones a nivel de frase

El modelo de Redes Semánticas se ha utilizado para representar una oración declarativa (Figura 1) que describa los distintos aspectos de un evento en concreto. (Rich & Knight, 1994).

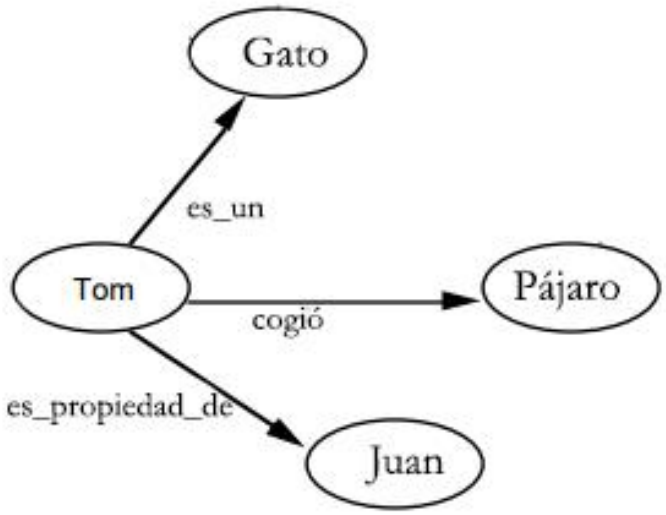

Figura 1. Red semántica de la oración "Tom, el gato de Juan, cogió un pájaro." (Huntbach, 1996)

Como se puede observar, es una buena alternativa a la Lógica ya con este modelo se puede representar la información que podría extraer un ser humano al leer esta oración como el tiempo pasado que denota el uso del pretérito perfecto simple del verbo *coger*.

4.2.2. Representaciones a nivel discursivo local

La representación de Redes Semánticas a nivel discursivo local se realizará uniendo, mediante el Principio de Composicionalidad, las distintas representaciones de las oraciones que compongan el texto corto en un único grafo (Figura 2).

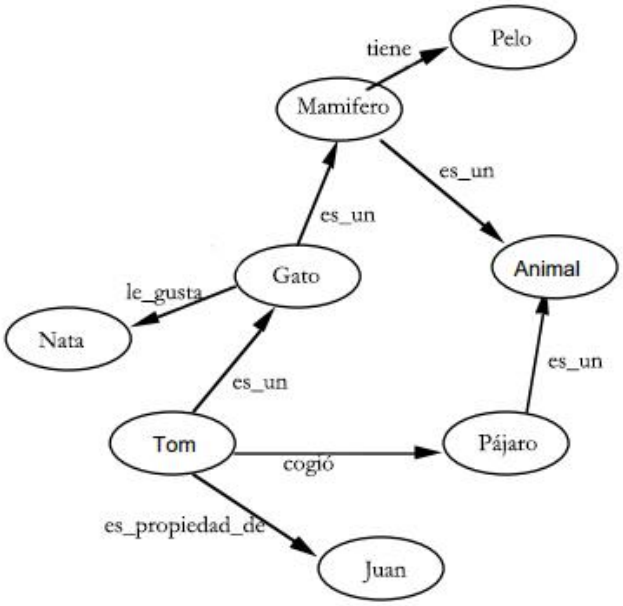

Figura 2. Red semántica del texto "Tom, el gato de Juan, cogió un pájaro. A los gatos les gusta la nata. Los gatos son mamíferos y todos los mamíferos tienen pelo y son animales. Los pájaros también son animales." (Huntbach, 1996)

No obstante, como se ha mencionado previamente, este tipo de representaciones son eficaces a nivel de frase o a nivel discursivo local pero, al aumentar la información que se desea representar, las redes serán cada vez más complejos hasta alcanzar un punto en el que son poco manejables y difíciles de interpretar.

4.3. Marcos

Desarrollado por Minsky (Minsky, 1974), el concepto de Marco deriva originalmente de las Redes Semánticas. Una estructura de Marcos es una ampliación de la estructura de Redes Semánticas ya que sustituye los nodos de este modelo por grupos estructurados de información. Cada Marco consiste en una colección de atributos, habitualmente llamados ranuras, con valores asociados y que describe alguna entidad o concepto del mundo. Están

organizados en dos niveles: el nivel superior contiene una información fija que siempre es cierta mientras que el nivel inferior se compone de ranuras que se rellenan mediante valores asociados que describen el elemento del nivel superior (Figura 3).

MARCO	Gato	← Nivel superior
NOMBRE	Dylan	← Nivel inferior
RAZA	Persa	
DUEÑO	Raúl	

Figura 3. Ejemplo de Marco.

4.3.1. Representaciones a nivel de frase

Para realizar una representación mediante el modelo de Marcos, se construyen sistemas de Marcos a partir de colecciones de estos y se conectan unos con otros, ya que puede darse que el valor de un atributo de un Marco puede ser, a su vez, otro Marco (Rich & Knight, 1994). Los Marcos, al estar conceptualmente conectados, permiten que los atributos de las entidades puedan ser heredados de otras entidades anteriores en la jerarquía.

Se utilizará el ejemplo de la Figura 4 para ilustrar.

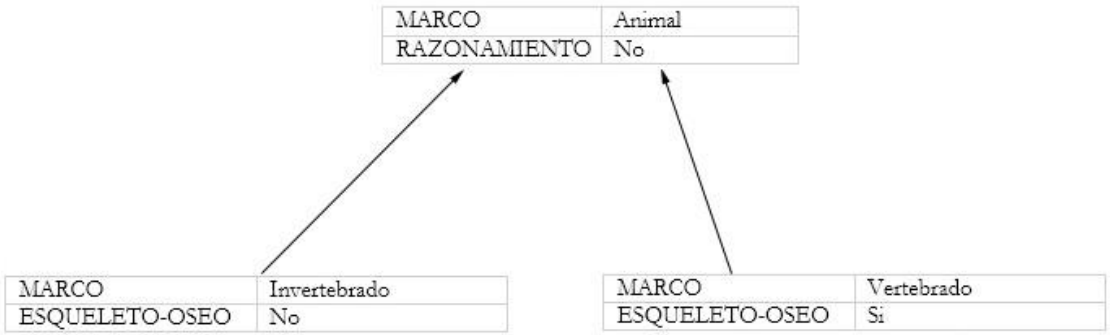

Figura 4. Marco de la oración "Los animales son irracionales y se clasifican en vertebrados e invertebrados." (Bernal Zamora, Frames)

Como se puede observar, este modelo es muy parecido al modelo de Redes Semánticas pero de forma más compacta es capaz de representar más información.

4.3.2. Representaciones a nivel discursivo local

Para utilizar este modelo para la representación a nivel discursivo, es necesario aplicar, de nuevo, el Principio de Composicionalidad. Por lo tanto, la representación final se realizará al unir las representaciones de las oraciones que compongan el texto.

Se ilustrará con un ejemplo (Bernal Zamora, Frames) y se representará en la Figura 5.

> "Los animales son irracionales y se clasifican en vertebrados e invertebrados. Los vertebrados poseen componente óseo y los invertebrados no. Los animales tienen movimiento propio y tienen nivel de inteligencia inferior. Los mamíferos son vertebrados con reproducción vivípara, sobreviven en tierra y tienen la piel cubierta de pelo. Los reptiles son vertebrados con reproducción vivípara, sobreviven en tierra y tienen la piel cubierta de pelo. Los reptiles son vertebrados de vida terrestre, cuerpo con escamas y se arrastran. Las aves son vertebradas, vuelan y son cuerpos cubiertos de plumas. Los peces son vertebrados con respiración branquial, cuerpo con escamas y medio de vida el agua. Los insectos son invertebrados, vuelan y su medio es terrestre. La ballena es un mamífero con piel lisa y vive en el agua."

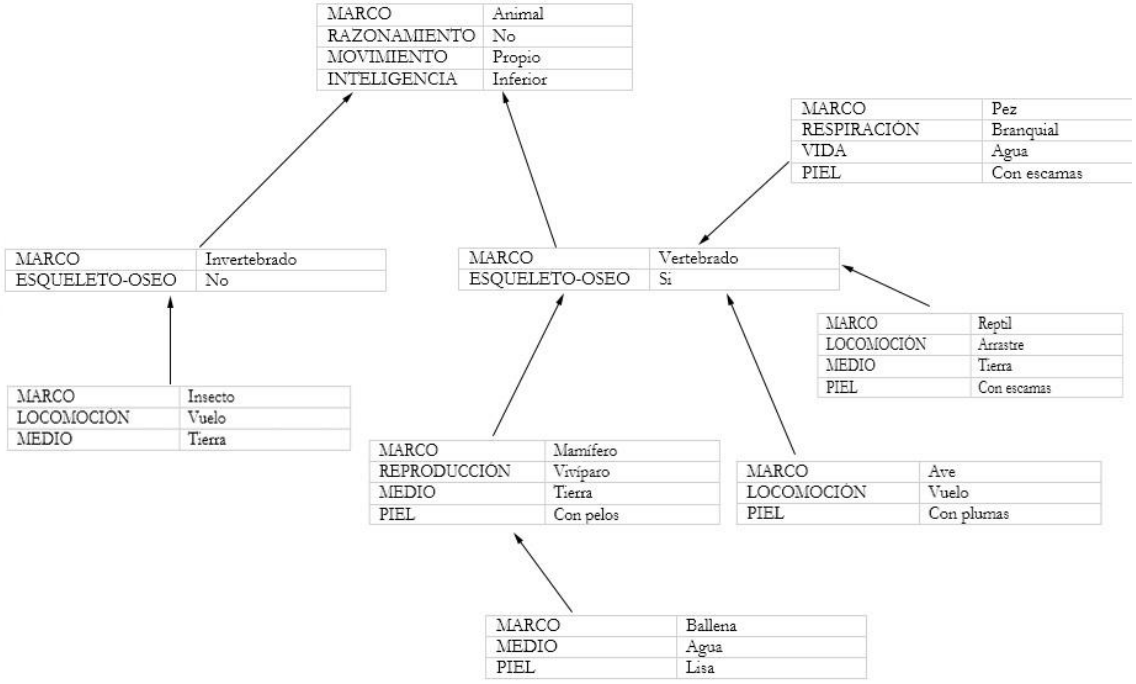

Figura 5. Representación de un pequeño texto usando el sistema de Marcos. (Bernal Zamora, Frames)

En definitiva, este modelo tiene mayor poder expresivo gracias a que las ranuras se pueden rellenar mediante diferentes valores que describen el elemento principal del Marco. Al representarse mediante Marcos relacionados unos con otros, es un modelo fácil de entender. No obstante, vuelve a tener la misma limitación que ya se vio con las representaciones lógicas y las Redes Semánticas. Son modelos eficaces a nivel de palabra o nivel discursivo local pero más allá resulta ineficiente de manejar y complejo de representar grandes cantidades de información.

4.4. Modelos de representación a nivel discursivo global

Los modelos de representación revisados en las secciones anteriores, modelos basados en la Lógica, en Redes Semánticas y en Marcos, pueden considerarse eficaces para textos cortos como son una frase o un párrafo. Para ello, se ha utilizado el Principio de Composicionalidad para ir construyendo la representación global del texto mediante la composición frase a frase. Sin embargo, si se aumenta la complejidad y se pretende hacer la representación a nivel discursivo global, las representaciones serán demasiado extensas, poco manejables y, en consecuencia, poco eficientes.

Por consiguiente, surgen otros modelos, como los que se presentarán a continuación. Estos modelos son 'masivos', es decir, son capaces de manejar una cantidad de texto impensable para los modelos vistos hasta ahora y funcionan de manera apropiada para la representación a nivel discursivo global.

4.5. Representaciones en Árbol y Segmentos

Este modelo es una técnica usada para representar texto de tamaño medio o grande, es decir, es un modelo usado para representar texto a nivel discursivo global. Los textos de este tamaño no se pueden ver como una secuencia lineal de oraciones, más bien se han de ver como oraciones que se agrupan en unidades, llamadas segmentos, que poseen una estructura jerárquica.

En el siguiente ejemplo (Allen, 1995) se puede encontrar un texto, en este caso una conversación entre dos personas. Fernando está ayudando a montar el cortacésped a Andrea.

"Fernando: Vale, tienes el motor terminado

Ahora conecta la cuerda a la parte superior del motor

Por cierto, ¿has comprado gasolina hoy?

Andrea: Sí, la cogí cuando compré la nueva rueda del cortacésped

Olvidé coger mi lata de gas, así que he comprado una nueva

Fernando: ¿Ha costado mucho?

Andrea: No, y podría usar otra de todas formas para mantener el tractor

Fernando: Ok

Andrea: ¿La has atado ya?"

El texto en este caso es en realidad un conjunto de frases que se unen en segmentos, los cuales tienen una estructura jerárquica como se ha mencionado previamente. Se define segmento como una secuencia de cláusulas con coherencia local (Allen, 1995). Existen dos enfoques que caracterizan, a su vez, lo que define un segmento. Por una parte, el enfoque intencional, que propone que todas las frases de un segmento contribuyen a un propósito común de todo el discurso. Por otra parte, el enfoque informativo, en el cual todas las frases de un segmento están relacionadas entre sí por alguna relación temporal, causal o retórica.

También hay que tener en cuenta los sub-segmentos que pueden surgir dentro de cada segmento. Normalmente, una situación discursiva tendrá muchos segmentos con sub-segmentos como se puede observar en la Figura 6.

SEGMENTO 1

a) Alberto y Pablo fueron a comprar un nuevo cortacésped
b) ya que el que tenían fue robado.

SEGMENTO 2

c) Pablo vio a los hombres que lo robaron y
d) los persiguió por la calle,
e) pero ellos se escaparon en un camión

f) Después de mirar en la tienda
g) se dieron cuenta de que no podían permitirse comprar uno nuevo

SEGMENTO 3

h) Por cierto, Alberto perdió su trabajo el mes pasado
i) por lo tanto últimamente ha tenido problemas de dinero.
j) Ha estado buscando un nuevo
k) pero por ahora no ha tenido nada de suerte

l) En fin, al final encontraron uno usado en un mercadillo

Figura 6. Representación en segmentos de una situación discursiva. (Allen, 1995)

Esta estructura se puede representar mediante un Árbol como se muestra en la Figura 7.

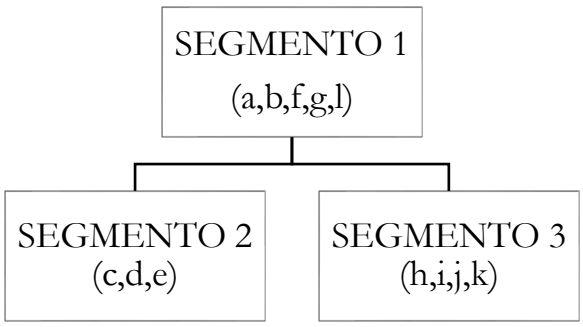

Figura 7. Representación de segmentos mediante un Árbol de una situación discursiva. (Allen, 1995)

En resumen, cada segmento muestra coherencia local, es decir, todas las cláusulas que componen cada segmento poseen un sub-tema que al organizarse jerárquicamente, constituye el tema principal del texto. Se elige la representación en Árbol porque es un tipo de representación fácil de comprender y visualizar.

4.6. Bolsa de Palabras

Este modelo y el siguiente, el modelo de Análisis Semántico Latente, provienen de la Informática, concretamente del área de Aprendizaje Automático. Los modelos de representación que se han presentado hasta este momento poseen una limitación muy importante, no "aprenden" de manera autónoma como lo harían los humanos. Es por este motivo por el que surgen los sistemas basados en Aprendizaje Automático los cuales superan esa limitación. Estos sistemas necesitan trabajar con grandes cantidades de información ya que, de lo contrario, no son eficaces.

El modelo de Bolsa de Palabras (Serrano Moreno, 2007) utiliza todas las palabras de un texto para representar la "semántica" del texto. Así las palabras se consideran las características del texto y, por lo tanto, la dimensión del "espacio semántico" de un texto será igual al número de palabras de dicho texto. Se ilustrará con los siguientes ejemplos:

a) 'A Miguel le gusta tocar la guitarra y a Marcos también le gusta.'

b) 'A Miguel también le gusta tocar el piano.'

Al utilizar todas las palabras como características, la Bolsa de Palabras se construye a partir de la frecuencia de aparición de los términos (TF). Por lo tanto, la Bolsa de Palabras estará compuesta por todas las palabras de las dos frases y los números corresponderán a la frecuencia de aparición. Como se puede observar en la Figura 8, en la primera frase, el término 'a' aparece dos veces, el término 'Miguel', una… etc.

Bolsa de Palabras	a	Miguel	le	gusta	tocar	la	guitarra	y	Marcos	también	el	piano
Frase a)	2	1	2	2	1	1	1	1	1	1	0	0
Frase b)	1	1	1	1	0	0	0	0	0	1	1	1

Figura 8. Representación de frases mediante sus bolsas de palabras.

Como es de esperar, las palabras más frecuentes serán las conocidas como *stop words*, es decir, palabras que no tienen ningún sentido léxico relevante para el texto (artículos, demostrativos, etc.) y es necesario quitarlas.

Además de obtener la frecuencia de aparición de los términos (TF), también se puede obtener el producto de la frecuencia de aparición de un término en un texto (TF) y su frecuencia inversa de texto. Es decir, cuanto mayor sea la cantidad de textos y la frecuencia de apariciones de un término, mayor será este factor, llamado IDF. Esto se lleva a cabo mediante el TF-IDF (Term-Frequency, Inverse document frequency).

Si se calcula el TF-IDF de 4 términos en distintos textos (Figura 9), los pesos que se obtienen denotan la importancia de cada término en cada texto (Figura 10).

TF (Frecuencia de aparición de cada término)			
Término	**Texto1**	**Texto2**	**Texto3**
biblioteca	27	4	24
Archivo	3	33	0
documento	14	0	17
Museo	0	33	29

Figura 9. Cálculo del TF-IDF.

Cálculo de pesos TF-IDF			
Biblioteca	TF-IDF (biblioteca,Texto1)	TF-IDF (biblioteca,Texto2)	TF-IDF (biblioteca,Texto3)
	27 x 2,65 = 71,55	4 x 2,65 = 10,60	24 x 2,65 = 63,60
Archivo	TF-IDF (archivo,Texto1)	TF-IDF (archivo,Texto2)	TF-IDF (archivo,Texto3)
	3 x 3,08 = 9x24	33 x 3,08 = 101,64	0 x 3,08 = 0
documento	TF-IDF (documento,Texto1)	TF-IDF (documento,Texto2)	TF-IDF (documento,Texto3)
	14 x 2,50 = 35	0 x 2,50 = 0	17 x 2,50 = 42,50
Museo	TF-IDF (museo,Texto1)	TF-IDF (museo,Texto2)	TF-IDF (museo,Texto3)
	0 x 2,62 = 0	33 x 2,62 = 86,46	29 x 2,62 = 75,98

Figura 10. Pesos de cada término.

El cálculo del TF-IDF también sirve para comparar la similitud de dos textos (por ejemplo, en cuanto a la temática) mediante una medida de cercanía. Esta medida puede ser la distancia Euclídea, que calcula la distancia entre dos puntos como la resta de sus componentes (los pesos de las palabas). No obstante, esta medida puede ser poco indicativa en el caso de textos con bolsas de palabras con pesos no normalizados, por lo que en este trabajo se utilizará la medida del coseno que calcula el ángulo entre las "bolsas de palabras". Con esta medida se obtienen valores comprendidos entre 0 y 1, de forma que, cuánto más cercano a 1 sea el número obtenido, mayor similitud tendrán los textos.

Si lo que se pretende es calcular es la similitud entre el Texto 1 y el Texto 2 y siendo x_1 la frecuencia de aparición del primer término del Texto 1 y x_2 la del segundo Texto, se realizará la siguiente operación:

$$\frac{x_1 \cdot y_1 + x_2 \cdot y_2 + \ x_n \cdot y_n}{\sqrt{x_{1^2} + \ x_{2^2} + x_{n^2}} \cdot \sqrt{y_{1^2} + \ y_{2^2} + y_{n^2}}} = 0{,}20$$

El resultado es 0,20. Es decir, los textos 1 y 2 no son muy parecidos.

Sin embargo, aunque el modelo es eficaz respecto al cálculo de semejanza entre textos tiene el problema de que no siempre las palabras que incluye el texto son representativas del tópico del texto. Por ejemplo, considérese un texto que habla de ciertas razas de perros. Es probable que no contenga la palabra "perro" aunque su tópico global es, ciertamente, "razas de

perros". Es, pues, necesario buscar otros modelos que sean capaces de captar mejor la semántica de un texto. El modelo más paradigmático de este siguiente nivel de representación semántica es el Análisis Semántico Latente.

4.7. Análisis Semántico Latente (LSA)

El modelo que se presenta a continuación es el Análisis Semántico Latente cuyas siglas en inglés son LSA y corresponden a Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998). La principal diferencia con el modelo de Bolsa de Palabras es que hace corresponder un conjunto de términos con un conjunto de conceptos. Este modelo, por lo tanto, utiliza como formalismo de representación textual la Bolsa de Conceptos en vez de la Bolsa de Palabras. La Bolsa de Conceptos (Sahlgren & Cöster, 2004) deja de entender los textos como una colección de palabras y comienza a tratarlos como una colección de conceptos. Esto se basa en la suposición de que el significado de un texto puede obtenerse al combinar los significados de sus términos y estos se representan en un espacio vectorial "semántico".

Se ilustrará con el clásico ejemplo propuesto por los autores de este modelo (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990).

Se propone una colección de nueve textos (Figura 11) en los que cada uno contiene un título y, a su vez, están divididos en dos grupos: el grupo 'C' trata sobre la interacción de hombres y máquinas y el grupo 'M', sobre gráficos.

C1	human machine interface for Lab ABC computer applications
C2	a survey of user opinion of computer system response time
C3	the EPS user interface management system
C4	system and human system engineering testing of EPS
C5	relation of user-perceived response time to error measurement
M1	the generation of random, binary, unordered trees
M2	the intersection graph of paths in trees
M3	graph minors IV: Widths of trees and well-quasi-ordering
M4	graph minors: A survey

Figura 11. Ejemplos propuestos por los autores. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

El primer paso es realizar una matriz con la frecuencia de aparición de cada concepto, donde cada celda indica el número de veces que un término aparece en un texto. Esta matriz se denominará matriz X (Figura 12).

Términos	Textos								
	C1	**C2**	**C3**	**C4**	**C5**	**M1**	**M2**	**M3**	**M4**
Human	1	0	0	1	0	0	0	0	0
Interface	1	0	1	0	0	0	0	0	0
Computer	1	1	0	0	0	0	0	0	0
User	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
Response	0	1	0	0	1	0	0	0	0
Time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
Survey	0	1	0	0	0	0	0	0	1
Trees	0	0	0	0	0	1	1	1	0
graphs	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Figura 12. Matriz X. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

Para obtener el verdadero Análisis Semántico Latente a partir de esta matriz, cada valor de la matriz deberá ponderarse por una función que expresa tanto la importancia de la palabra en el contexto como el grado de información que esa palabra contiene del dominio del discurso en general. Esta matriz se descompone mediante la descomposición en valores singulares (SVD). Esta descomposición es un método matemático el cual descompone una matriz en un producto de otras tres matrices (U, S y V). La matriz S es una matriz diagonal cuyos valores en la diagonal son distintos a cero. Estos valores son los valores singulares o característicos de la matriz original. En el caso de textos y términos los valores singulares de la matriz S diagonal se interpretan como los conceptos básicos respecto de los cuales se refiere cada texto. Esta reducción no sólo permite manejar mejor la matriz de términos al no ser tan grande como la original, ya que no es viable tener un espacio de, por ejemplo, 80.000 términos, sino que obtiene el "espacio semántico vectorial" en el que términos y conceptos están representados por medio de vectores que contienen solo la información sustancial para la formación de conceptos.

Por lo tanto, la matriz X se descompondrá en el producto de 3 matrices (Figura 13). La matriz S es la matriz diagonal que describe los valores singulares que son los conceptos principales del espacio semántico del conjunto de textos (Figura 15). Cada valor que no es cero

representa un concepto con un determinado peso. Los textos se representan como una combinación lineal de los conceptos principales.

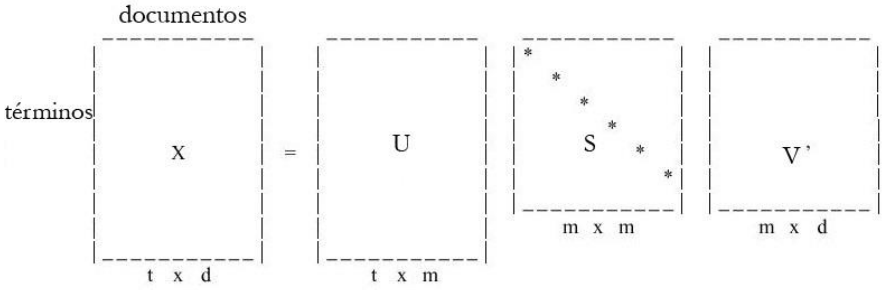

Figura 13. Descomposición de la matriz X. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

U =

0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18

Figura 14. Matriz U. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

S =

3.34								
	2.54							
		2.35						
			1.64					
				1.50				
					1.31			
						0.85		
							0.56	
								0.36

Figura 15. Matriz S. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

V =

0.20	-0.06	0.11	-0.95	0.05	-0.08	0.18	-0.01	-0.06
0.61	0.17	-0.50	-0.03	-0.21	-0.26	-0.43	0.05	0.24
0.46	-0.13	0.21	0.04	0.38	0.72	-0.24	0.01	0.02
0.54	-0.23	0.57	0.27	-0.21	-0.37	0.26	-0.02	-0.08
0.28	0.11	-0.51	0.15	0.33	0.03	0.67	-0.06	-0.26
0.00	0.19	0.10	0.02	0.39	-0.30	-0.34	0.45	-0.62
0.01	0.44	0.19	0.02	0.35	-0.21	-0.15	-0.76	0.02
0.02	0.62	0.25	0.01	0.15	0.00	0.25	0.45	0.52
0.08	0.53	0.08	-0.03	-0.60	0.36	0.04	-0.07	-0.45

Figura 16. Matriz V. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

Para lograr que la matriz sea más manejable, los autores proponen aislar los dos primeros valores singulares tanto de la matriz U, S y V y llamarlos, respectivamente U', S' y V' (Figura 17).

X≈ U′ S′ V′

0.22	-0.11	3.34		0.20	0.61	0.46	0.54	0.28	0.00	0.02	0.02	0.08
0.20	-0.07		2.54	-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.24	0.04											
0.40	0.06											
0.64	-0.17											
0.27	0.11											
0.27	0.11											
0.30	-0.14											
0.21	0.27											
0.01	0.49											
0.04	0.62											
0.03	0.45											

Figura 17. Matrices U', S' y V'. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

El último paso es multiplicar las tres matrices para obtener la matriz reducida que es la que los autores proponen como idónea para trabajar con ella. Esta matriz se denomina Xhihat (Figura 18).

Xhihat =

0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Figura 18. Matriz Xhihat. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

Esta nueva matriz permite comparar dos textos al obtener el producto escalar (que da la medida del coseno del ángulo) entre dos columnas de dicha matriz. Esto indicará la similitud entre dos textos a través de sus términos. También se pueden comparar dos términos al hacer lo mismo pero entre filas. Si lo que se pretende es comparar un término con un texto, es necesario fijarse en que cada celda de esta matriz relaciona un texto con un término.

En la representación gráfica de esta matriz (Figura 19), los términos se representan como puntos negros y los textos como cuadrados blancos. Las filas se interpretan como coordenadas de los puntos que representan los textos. Si se realiza el cálculo del coseno del ángulo entre dos puntos, se obtendrá la similitud entre ellos.

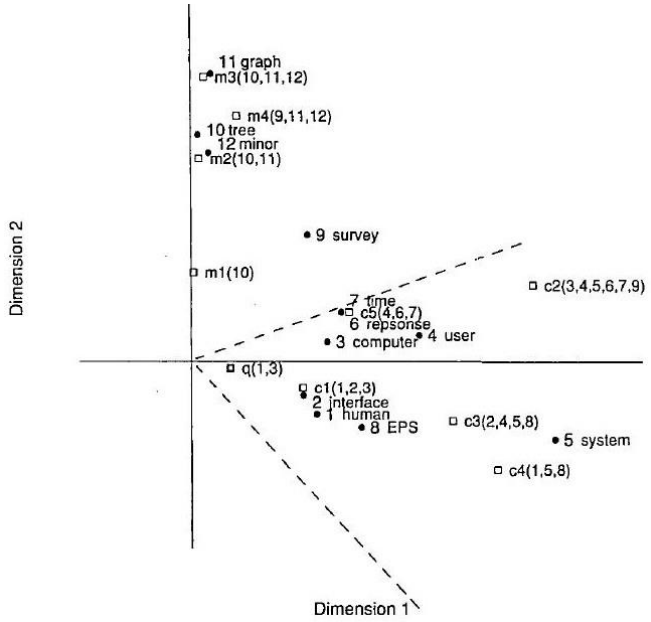

Figura 19. Representación gráfica de la matriz Xhihat. (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)

Sin embargo, estos procesos requieren una cantidad considerable de tiempo ya que puede darse el caso de tratar con millones de palabras, lo que quiere decir, millones de filas en las matrices. Es un modelo muy costoso en tiempo y memoria del sistema informático que calcule y manipule estas representaciones (Serrano Moreno, 2007).

5. Resumen, conclusiones y trabajo futuro

En este trabajo se han revisado seis modelos de representación textual usados a lo largo de la historia de la Lingüística Computacional para resolver la problemática aquí planteada: la representación del significado intrínseco del texto.

Se ha comenzado por realizar una introducción al problema, situándolo y justificando la elección de texto como objeto de este trabajo, apuntando su interés científico y económico.

A continuación se ha expuesto la caracterización del texto, señalando a la coherencia como el elemento más importante para conocer cómo se obtiene el significado intrínseco del texto. La coherencia entre las frases que constituyen el texto viene determinada por el Marco de Integración Global. Se ha aplicado esta idea al Principio de Composicionalidad para emplearlo en la representación del significado intrínseco del texto mediante los modelos presentados en el trabajo.

Se han considerado los modelos más representativos y paradigmáticos que resuelven la problemática según el nivel que son capaces de representar. Se ha demostrado que los modelos de representación del significado basados en Lógica, Redes Semánticas y Marcos son eficaces a la hora de realizar representaciones de textos cortos. Representan de manera eficiente frases y aplicando el Principio de Composicionalidad son capaces de representar textos cortos a nivel discursivo local de forma eficiente. Sin embargo, al aplicarlos para representar texto de tamaño medio y grande, a nivel discursivo global, estos modelos no son una buena solución, pues el tamaño de las representaciones finales es demasiado extenso como para ser fácilmente manejable y entendible.

Para representar textos de tamaño medio y grande a nivel discursivo global, se han señalado tres modelos que sí son eficaces a la hora de realizar esta tarea: Árbol y Segmentos, Bolsa de Palabras y Análisis Semántico Latente. Estos modelos, sobre todo los dos últimos, al trabajar con una cantidad ingente de información, necesitan hacer uso de importantes recursos informáticos, fundamentalmente de memoria y procesador, por lo que pueden resultar demasiado costosos. Sin embargo, actualmente son los que ofrecen resultados aceptables en aplicaciones comerciales de procesamiento del lenguaje natural como son los clasificadores de texto o los sistemas de recuperación de información.

Debido a la extensión limitada del Trabajo Fin de Grado, no se ha podido realizar una revisión completa de los modelos de representación semántica del texto. Quedan, sin embargo, las fuentes bibliográficas recogidas en la bibliografía con las que se puede seguir ampliando la revisión en posibles trabajos futuros.

6. Bibliografía

Allen, J. (1995). *Natural language understanding* (2nd ed). Redwood City, Calif: Benjamin/Cummings Pub. Co.

Bernal Zamora, Leonardo (s. f.). *Frames*. Recuperado a partir de https://es.slideshare.net/leonardobernalzamora/frames-5316032

Bernárdez, E. (1982). *Introducción a la lingüística del texto*. Espasa-Calpe.

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, *9*(2), 48-57.

Carreter, F. L. (1971). *Diccionario de términos filológicos*. Editorial Gredos.

Cintula, P., Jek, P. H., & Noguera, C. (Eds.). (2011). *Handbook of Mathematical Fuzzy Logic. Volume 1*. London: College Publications.

Clocksin, W., & Mellish, C. S. (2003). *Programming in Prolog*. Springer Science & Business Media.

Collin, & Quillian. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*.

Cuenca, M. J. (2010). *Grámatica del texto*. Madrid: Arco Libros.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Fausett, L. (1994). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications* (Edición: 01). Englewood Cliffs, NJ: Financial Times Prentice Hall.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge ; New York: Cambridge University Press.

Finlayson, M. A., Richards, W., & Winston, P. H. (2010). Computational Models of Narrative: Review of a Workshop. *Mark Finlayson*.

Frege, G. (1948). Sense and Reference. *The Philosophical Review*.

Garson, J. W. (2013). *Modal Logic for Philosophers* (2 edition). New York: Cambridge University Press.

Grupo PM. (2001). Enciclopedia Canina. Recuperado a partir de http://www.mascotaspfi.com/descargas/perros.pdf

Hilbert, M. (2014). What Is the Content of the World's Technologically Mediated Information and Communication Capacity: How Much Text, Image, Audio, and Video? *The Information Society*, *30*(2), 127-143.

Huntbach, M. (1996). Artificial Intelligence I. Notes on Semantic Nets and Frames.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-84.

Minsky, M. (1974). *A Framework for Representing Knowledge*. Technical Report. Massachusetts Institute of Technology, Cambridge, MA, USA.

Rich, E., & Knight, K. (1998). *Inteligencia artificial.* (P. A. González Calero & F. Trescastro Bodega, Trads.). Madrid

Sahlgren, M., & Cöster, R. (2004). Using Bag-of-concepts to Improve the Performance of Support Vector Machines in Text Categorization. En *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Serrano Moreno, J. I. (2007). *Modelo computacional de lectura cognitiva para la representación automática de textos*. Madrid: Universidad Complutense de Madrid, Servicio de Publicaciones.

Yao, M. (2017). *Conversational Interfaces: Principles of Successful Bots, Chatbots, Messaging Apps, and Voice Experiences*. Topbots.