Una Representación Alternativa para Textos

Daniela López De Luise*

I. Introducción

Hasta la actualidad, el manejo de grandes colecciones de datos ha presentado complejas soluciones, algunas más competentes que otras. Entre las estrategias consideradas pueden mencionarse la reducción de dimensiones [4], text mining[3], machine learning [5], variantes de Natural Language Processing[6], etc. La presente es una propuesta que pretende hallar una solución reorganizando sistemáticamente los contenidos textuales o no. Es posible utilizar la misma para ver en forma simbólica y esquemática un texto y a la vez representar metódicamente su contenido. Esto serviría de base para estructuras de procesamiento automático como la presentada en [1], para las estructuras representativas de oraciones textuales (allí denominadas Estructuras de Composición Interna o E_{ci}).

Se presentará una mecánica austera y pragmática, ya probada empíricamente en [1], donde se reprodujo un diagrama resultante de la transformación que aquí se describe como parte de una encuesta. En ese mismo trabajo, como resultado de la mencionada encuesta, se halló que la totalidad de los encuestados fue capaz de contestar preguntas acertadamente a partir del contenido simbolizado.

El resto de este trabajo se organiza de la siguiente manera: sección II descripción de los fundamentos, sección III, un caso de muestra, sección IV casos de ambigüedad, sección V contradicciones y errores. Finalmente en la sección VI se darán algunas conclusiones y trabajo a futuro.

II. Descripción de los Fundamentos

El objetivo de esta propuesta es generar una red de objetos automáticamente y utilizar la red como una organización de los contenidos textuales o no para facilitar el uso de los mismos (por caso en [1] serviría para procesar los denominados E_{ce}). La red será una objetivación de las relaciones entre palabras adyacentes, utilizando para ello simplemente el significado generalmente más sencillo y común de ciertas palabras especiales aquí llamadas ***conectoras.*** De este modo se organizaría la búsqueda automática del locus (zona dentro de la red donde se halla el contenido deseado).

Los principios de esta organización son sencillos:

- Los sustantivos (singulares, plurales, colectivos, etc) existen en el texto y son identificables.
- Las palabras se conectan con otras palabras o símbolos definidos (artículos, puntuación, etc.) y acotados en cantidad. A éstos se los denomina ***conectores***.

* Docente de la Facultad de Ingeniería - UP y Directora del ITLab.

- Los conectores tienen una función bien definida en el texto.
- Algunos conectores son prescindibles para la estructura simbólica y por lo tanto pueden eliminarse de la misma. A estos se los denomina ***conectores de eliminación***.
- Las estructuras E_{ci}. No son rígidas sino que pueden variar con la probabilidad de su aparición y según el sesgo del aprendizaje; en otras palabras el objetivo de construcción del E_{ci}, O_{eci}.

La tabla 1 muestra un compendio de transformaciones básicas que se proponen para construir una E_{ci}. Para comprender cómo se maneja tomemos por ejemplo, una porción de texto donde se encuentra lo siguiente: "....apertura de puertas y ventanas.....".

Es claro que esta estructura se puede asimilar al caso 1 de la tabla, tomando "apertura" como A, "puertas" como B y "ventanas" como C. En consecuencia debiera aplicarse la representación simbólica que figura en la tercer columna, quedando algo como lo que se muestra en la Fig. 1. Pueden notarse que algunos casos especiales:

a) Hay casos que tienen en su descripción una frase del estilo "aplica marcador ⏏". Esto significa que por su colocación en la frase, alguna de las palabras involucradas son candidatas a representar toda la frase.
b) Otra descripción interesante es la que indica el caso 6: "invierte la secuencia de inserción hasta el próximo sustantivo", lo que indica que a partir de allí cada nuevo caso se inserta en orden inverso, es decir, invirtiendo la flecha direccional que conecta las palabras en la estructura de la E_{ci}.
c) Algunos casos provocan un cambio en el lugar donde se continuará insertando dentro de la red de símbolos (Ej. El caso 8).
d) En el caso 23 se indica que los subesquemas que se estén desarrollando se suspenden. Esto significa por ejemplo que si se está procesando un caso 7, ya no se le sigue procesando.
e) en general los casos se aplican recursivamente. Es decir que si al procesar una posición B aparece otro caso, primero se resuelve el primer caso y luego el último en aparecer.
f) En todos los casos A, B, C pueden ser más de un sustantivo secuenciado, y se tomarán como un conjunto único.

Fig. 1: Representación simbólica de una porción de texto.

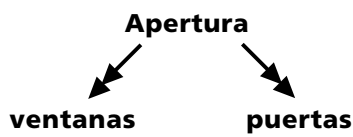
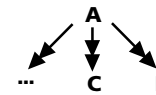
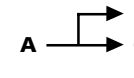
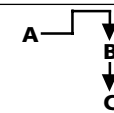
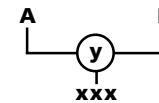
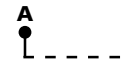
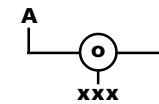
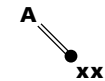
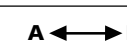
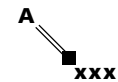
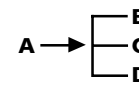


Tabla 1: Conversión a símbolos

ID	caso	símbolo
1	**A de B [y | o C[y |o ...]]** **A de B [, C[,...y | o...]]** Disgrega A en B, C,...Luego, si sólo existe B continúa la inserción en B, si existen C, D, etc. Continúa en A.	
2	**A para B [, C...]** Compone los "para" como selección	
3	**A con B** Hila conceptos / sustantivos	
4	**A y/ e B xxx** Junta una enumeración. Coordina el último sustantivo a izquierda con los sustantivos a derecha de "y" procesando lo que sigue como "xxx". Si xxx contiene más de una palabra, entonces saltea desde el primer sustantivo que halla hasta el próximo **"."**, **":"**, **","**, **";"** o fin de texto. Si en B aparecen los casos 1-3, 7-9, 11-15, 18,21, 24-31, entonces se tratan estos casos primero y se continúa en xxx	
5	**A, B** Coordina B al mismo nivel que el último sustantivo generando un caso 4. Si al retroceder en el texto halla un caso 3, deberá tomar el primer sustantivo que se presente antes de la ocurrencia del caso 3. Si queda como primer palabra después de punto, entonces ignora la coma y continúa insertando donde estaba. - Aplica marcador el marcador ▲ a B - aplica a los casos en que aparece ;:. Antes que otra coma	
6	**A por B** Subordina A al B. -Invierte la secuencia de inserción hasta el próximo sustantivo	
7	**A o B xxx** Si en B aparecen los casos 1-3, 7-9, 11-15, 18,21, 24-31, entonces se tratan estos casos primero y se continúa en xxx	

ID	caso	símbolo
8	**A en B[, C,...[y |o ...]]** Busca el próximo sustantivo anterior de la estructura E_{ci} y le marca con ▲ también le asocia con el símbolo especial a B ($\in$). Continúa en B la inserción. Si no existe un sustantivo o hay antes un caso 4 o 7,entonces elimina "B" y "en" y continúa la inserción donde estaba. Si en la búsqueda encuentra casos 1 o 3, los saltea y continúa la búsqueda hacia atrás.	**A $\in$ B[, C,...]**
9	▲ marca para futura entrada de rastreo o búsqueda	
10	**A: xxxx** delimita lo que sigue hasta el final de la frase (. ó ;) y subordina xxx. Luego sigue insertando en A	
11	**A Contra B**	
12	**A de su B** Es similar al caso 1 pero invierte el proceso de inserción de los sucesivos componentes.	
13	**"xxx"** (xxx) {xxx} ,xxx, similar al caso 10, pero el texto delimitado se lo subordina al sustantivo anterior. En caso de no existir se marca el primer sustantivo del contenido con ▲ y se genera una nueva E_{ci}. Luego la inserción continúa desde el punto A.	
14	**A a B** Enlaza A con B	
15	**A al B** Enlaza A con B	
16	**A entre las cuales B, C y D** Este caso se aplica sólo si A es sustantivo	
17	**,y** Elimina la "y" y actúa como si fuera un caso 5	
18	**A quien(es) B** Vincula B al anterior sustantivo. Similar al caso 14	
19	**A, para quien(es) B** Vincula B al anterior sustantivo. Similar al caso 18 pero invierte A con B	

ID	caso	símbolo
20	**A para quien(es) B** Similar al caso 2 pero se elimina B.	
21	**A que B**	**A** ⌈ **B**
22	**;** Busca el primer sustantivo desde el principio de la E_{ci} y retoma el proceso de inserción de lo que siga a partir de allí. Si no hay un sustantivo, toma la primer palabra que no sea un caso especial. Si no tiene éxito, genera una nueva E_{ci}.	
23	**.** Si luego sigue un sustantivo genera una nueva E_{ci}, de lo contrario: busca el primer sustantivo desde el principio de la E_{ci} y retoma el proceso de inserción de lo que siga a partir de allí. Si no hay un sustantivo, toma la primer palabra que no sea un caso especial. Si no tiene éxito, genera una nueva E_{ci}. Los subesquemas que se estén desarrollando se suspenden.	
24	**A ante B**	**A** —•→ **B**
25	**A según B**	**A** —÷→ **B**
26	**A sin B**	**A** —x→ **B**
27	**A so B**	**A** —S→ **B**
28	**A sobre B** Conecta A con B	**A** —–→ **B**
29	**A tras B**	**A** —=→ **B**
30	**A como B**	**A = B**
31	**A se B** No genera flecha direccional entre se y B	**A** → **B**
32	**A. Por esta razón B** **A. Por esto B** **A. Por eso B** **A Por lo tanto B** **A, Por esta razón B** **A, Por esto B** **A, Por eso B** **A, Por lo tanto B** **A, y Por esta razón B** **A, y Por esto B** **A, y Por eso B** **A, y Por lo tanto B**	**A** ⟹ **B**

Además de estos casos, se consideran los conectores de eliminación presentados en la tabla 2. Estos conectores sólo deben ser eliminados y no provocan ningún impacto dentro de la red de símbolos (E_{ci}).

Tabla 2: Conectores de eliminación

Conector de eliminación	
La	Ella
El	Ellas
Le	Les
Los	Ello
Las	Ellos
Los cuales	Un
Las cuales	Una
La cual	Esta
El cual	Este
Estas	Estos
Eso	Esos
Esa	Esas
Tan	Su
Sus	según*
Sin*	so*
sobre*	tras***
que	Además**
También***	Incluso***
Sin embargo	No obstante

Para el resto de los casos se toman directamente las palabras tal como se presentan en el texto y se reproducen unidas por una recta direccional desde la palabra anterior a la que le sigue. Gráficos, fotos, dibujos, sonidos, y cualquier otro elemento no textual no es procesado de esta manera sino reemplazado según la tabla 3.

* estos casos se aplican cuando inmediatamente antes hay un caso 13
** cuando están precedidos de caso 4 ó 7
*** cuando son primer palabra después de un punto

Algunas consideraciones que surgen de la tabla son especialmente aplicables a los textos bajados de Internet. Por ejemplo las palabras consideradas URL o Título:
URL: guarda el URL y lo marca con ▲ . Clasifica según el protocolo. Ej: *http:// www.democracia-diario.com.ar/2002/250702a.htm* genera algo como lo que se muestra en la Fig. 2. Cada sustantivo es marcado también.
TITULO: genera una marca en cada sustantivo. Genera un sustantivo TIT con la marca. Genera un sustantivo TXT sin la marca y lo asocia.

Fig. 2: Representación de un título

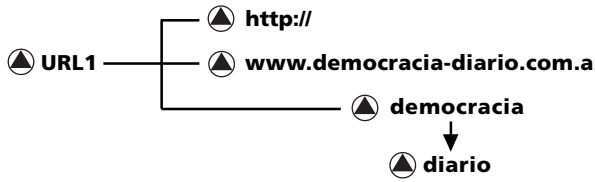


Tabla 3: Conectores especiales de textos Web

Objeto	Representación
URL	URLj. ▲ Genera un suevo E_{ci}
Link	LINK $_j$
Objetos wav	WAV $_j$
Objetos avi	AVI $_j$
Objetos jpg, jpeg	JPG $_j$
Objetos tiff	TIFF $_j$
Objetos doc	DOC $_j$
Objetos pdf	PDF $_j$
Objetos bmp	BMP $_j$
Tag <TITLE>	Sustantivo TIT, genera un nuevo E_{ci}.

También debe considerarse que cada nueva E_{ci} cuelga del sustantivo TXT si es que no puede colgar de otra frase relacionada por algún conector.

III. Caso de muestra

Para mostrar la aplicación de la representación se tomó un texto de una página Web. Luego se le aplicaron las normas de la representación. El texto procesado se muestra en el Apéndice y las redes de símbolos resultantes se muestran en las Fig. 3 y 4.

Fig. 3: E_{ci} de un caso

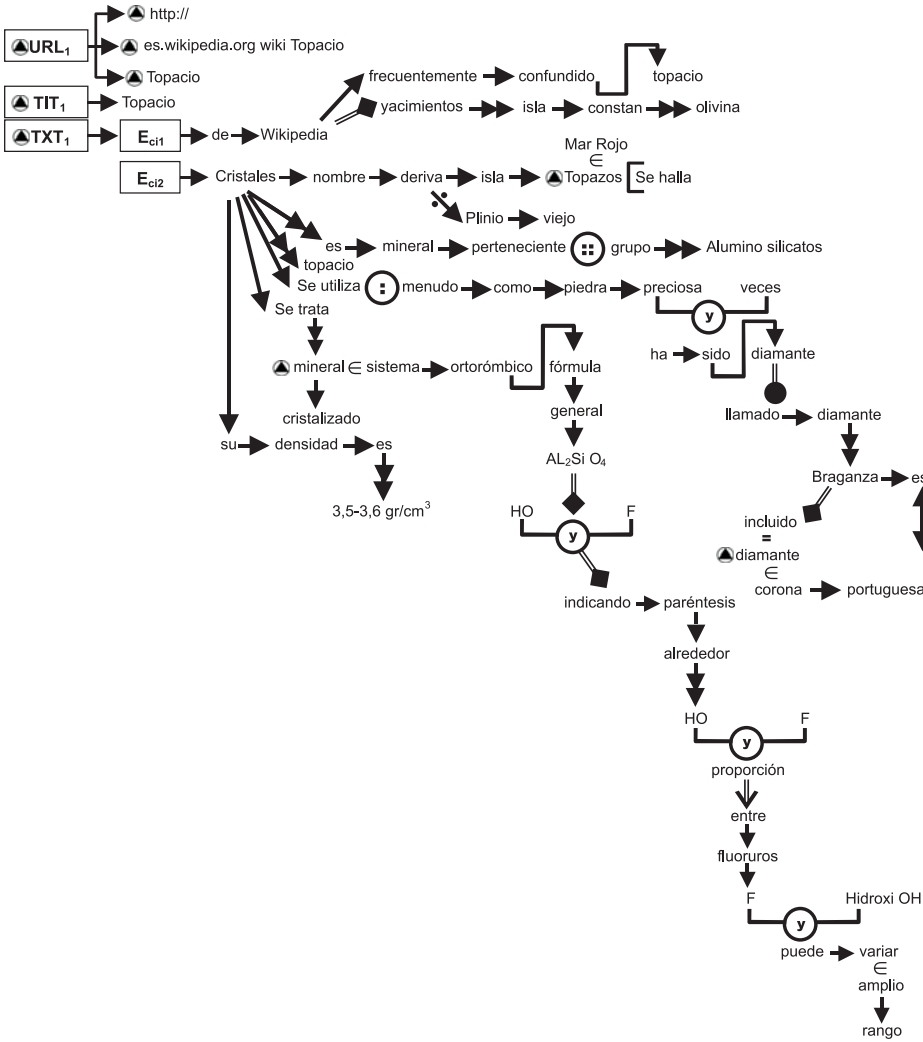
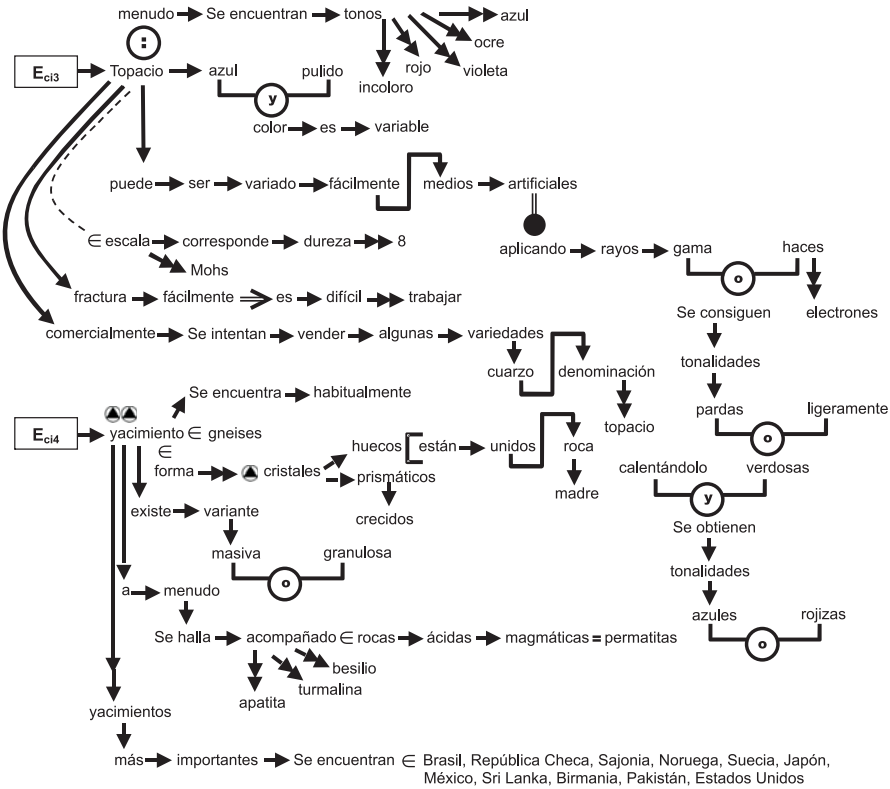


Fig. 4: E_{ci} de un caso (continuación).



IV. Manejo de ambigüedad

Es notable que las ambigüedades aquí tienen un significado simbólico con sólo extender adecuadamente la representación. Supóngase por caso la codificación de la tabla 4.

Tabla 4: Conectores de ambigüedades en textos

Palabra	Representación
Muy xxxx	{.7:xxxx}
Demasiado xxxx	{.1.1:xxxx}
Bastante xxxx	{.5:xxxx}
Algo xxxx	{.3:xxxx}
Poco xxxx	{.2:xxxx}
Casi nada xxxx	{.1:xxxx}

Entonces se tendrá la posibilidad de establecer un tipo de relación entre cierto texto preliminar y xxxx (en este caso la relación cuantifica a xxxx). Eventualmente, durante el procesamiento automático de una búsqueda, se podría considerar la valuación numérica sistemática como métrica de soporte para consideraciones especiales como:

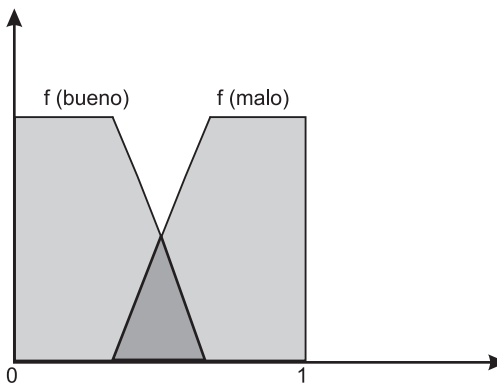
- si conviene o no continuar la búsqueda por esa rama de la red
- evaluación de importancia de la frase almacenada subsecuentemente
- búsquedas alternativas, etc.

V. Manejo de contradicciones

Las contradicciones semánticas son transparentes a este procesamiento del lenguaje y a lo sumo pueden presentarse palabras antónimas en lugares cercanos de la red. Lejos de ser una desventaja esto posibilita la búsqueda y procesamiento de las mismas como un caso más. Para ello será menester nuevamente extender el lenguaje a fin de facilitar la identificación de estos casos. Una posibilidad podría ser considerar un universo cerrado donde las alternativas de contradicción presentan una función de pertenencia. Supóngase por caso el siguiente tratamiento:

- se conserva una tabla de antónimos: blanco-negro, alto-bajo, lindo-feo, etc.
- se consideran automáticamente funciones de pertenencia como en la Fig. 5

Fig. 5: Funciones de pertenencia para antónimos



Para el procesamiento automático de estos casos la función serviría para consideraciones similares a las planteadas en la sección anterior para las ambigüedades:

-si conviene o no continuar la búsqueda por alguna de las ramas de la red solamente
-evaluación de importancia de la frase almacenada subsecuentemente
-búsquedas alternativas, etc.

Nótese que los casos en que las contradicciones incluyan más de dos alternativas, sólo será necesario definir las funciones de pertenencia adecuadamente y extender el tratamiento. Para evaluar la calidad de la representación se podría utilizar métricas e indicadores como en [2], garantizando de esta manera la representatividad de la solución propuesta.

VI. Conclusiones y trabajo a futuro

Se presentó un mecanismo de reorganización de contenidos que tendrá su justificación cuando sea incorporado a una estructura adecuada de procesamiento (como en [1]). Asimismo será necesario depurar algunos casos especiales de contradicciones y ambigüedades. Para un tratamiento más flexible y universal se estudiará la representación con XML.

II. Referencias

[1] M. D. López De Luise, "A Morphosyntactical Complementary Structure for Searching and Browsing". Proceedings of International Conference on Systems, Computing Sciences and Software Engineering (SCSS 2005). Dic. 2005. Bridegport University-IEEE.

[2] M. D. López De Luise, "Aplicación de Métricas Categóricas en Sistemas Difusos", Revista IEEE América Latina, marzo 2007.

[3] K. Lagus, "Text Mining with WEBSOM", Acta Polytechnica Scandinavica. Mathematics and Computer series No 110. 2000.

[4] S. Kaski, "Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering", Proceedings of the IJCNN. Int. joint Conference on Neural Networks.Vol 1.IEEE Service Center.Piscataway. NJ. pp 413-418. 1998.

[5] S. Finch, N. Chater, "Unsupervised methods for finding linguistic categories", Artificial Neural Networks 2. pages II-1365-1368.North-Holland. 1992.

[6] R. Miikkulainen, "Subsymbolic Natural Language Processing: An Integrated Model of Scripts, lexicon and memory", MIT Press, Cambridge MA. ISBN 0262132907. 1993.

Apéndice

Topacio
De Wikipedia

Cristales de topacio. El topacio es un mineral perteneciente al grupo de los aluminosilicatos. Su nombre deriva, según Plinio el Viejo, de la isla Topazos que se halla en el Mar Rojo. Sin embargo, los yacimientos de esta isla constan de olivina, frecuentemente confundida con el topacio.

Se utiliza a menudo como piedra preciosa y algunas veces ha sido confundido con el diamante: el llamado Diamante de Braganza, incluido como diamante en la corona portuguesa, es un topacio.

Se trata de un mineral cristalizado en el sistema ortorómbico con la fórmula general Al2SiO4(OH, F)2, indicando el paréntesis alrededor de HO y F que la proporción entre fluoruros F e hidroxi OH puede variar en un amplio rango, aunque su suma siempre será constante.

Su densidad es de 3,5 - 3,6 gr/cm³

Topacio azul, pulido el color es variable; a menudo se encuentran tonos de ocre, azul, violeta, rojo o incoloro. Además, puede ser variado fácilmente con medios artificiales: aplicando rayos gama o haces de electrones se consiguen tonalidades pardas o ligeramente verdosas y calentándolo se obtienen tonalidades azules o rojizas.

En la escala de Mohs le corresponde dureza de 8. Sin embargo, fractura fácilmente y por esta razón es difícil de trabajar.

Comercialmente se intentan vender algunas variedades de cuarzo con denominación de topacio.

Yacimientos

Se encuentra habitualmente en forma de cristales prismáticos crecidos en huecos que están unidos con la roca madre. Además existe una variante masiva o granulosa. A menudo se halla acompañado de berilio, turmalina y apatita en rocas ácidas magmáticas como las permatitas. También se encuentre en gneises.

Algunos de los yacimientos más importantes se encuentran en Brasil, República Checa, Sajonia, Noruega, Suecia, Japón, México, Sri Lanka, Birmania, Pakistan y los Estados Unidos.