Procesamiento del Lenguaje Natural, num. 29 (2002), pp.265-272 recibido 27-05-2002; aceptado 04-07-2002

Una gramática de estilos para resumir textos en español

Jacinto A. Dávila Quintero
Universidad de Los Andes. Centro de Investigación y Proyectos en Simulación y Modelos. CESIMO.
Mérida, Venezuela
5101
jacinto@ula.ve

Hilda Y. Contreras Zambrano
Universidad de Los Andes.
Postgrado en Computación.
Mérida, Venezuela
5101
hyelitza@ula.ve

Resumen: Este artículo presenta un experimento lingüístico que consiste en resumir textos escritos en español. El resumen es realizado con una herramienta computacional que aplica técnicas simbólicas basadas en una "gramática de estilos". Esta gramática modela las reglas de estilo para la escritura propuestas por Williams (1990). El programa puede obtener desde los tópicos de las oraciones de cada párrafo y reconocer elementos sintáctico-estructurales de cohesión y coherencia textual, hasta los tópicos más importantes del párrafo. Se aprovechan estos resultados para construir un resumen con oraciones asociadas a dichos tópicos. Esta versión inicial de nuestro resumidor muestra como en base a reglas lógicas para definir estilos y "tópicos" se pueden obtener resúmenes de textos "aceptables" por expertos en el dominio de conocimiento de los textos. También se sugiere que aplicando estos modelos podemos obtener resultados reduciendo la complejidad del procesamiento morfológico, sintáctico y semántico tradicional.
Palabras clave: Resumen Automático de textos, Lingüística Textual, Extracción de Información.

Abstract: This paper describes an experiment on text summarization. A summary is made by means of a logic program executed by a computer. The logic program is an embodiment of a symbolic technique for natural language processing based on "style grammars". These grammars, in turn, are based on a proposal by J. Williams (1990). The program obtains topics (the themes of the sentence, according to Williams) from each sentence in a paragraph and check its syntax and structure for cohesion and structural coherence. It ends with a proposal for the most important topic of the paragraph, which can be used as a building block for a summary. The main outcome of this work is the evidence that a set of rules, written in the language of logic, can embody style criteria, produce "topics" for texts in Spanish and lead to a tractable, computational implementation.
Keywords: Automatic Text Summarization, Text Linguistics, Information Extraction.

1 Introducción

Este artículo presenta un experimento con una técnica basada en manipulación simbólica para resumir textos en español explotando las características de cierto estilo de escritura. Para este ejercicio hemos usado las recomendaciones para una escritura clara, precisa y coherente dictadas por Williams (1990) y las hemos convertido en programas lógicos. Estos programas pueden reconocer características de estilo y las usan para extraer "tópicos" a partir de textos que usen (aún parcialmente) los elementos de estilo de Williams.

Decimos que un "tópico" es una frase que sugiere el tema del cual trata el texto. Esta definición que presentamos es, desde luego, vaga. Sin embargo, nos ha servido para vincular esta noción a lo que Williams llama tópico y con ello hemos usado su caracterización de tópico para producir un conjunto de cláusulas que lo definen. El resultado principal que

ISSN 1135-5948

reportamos aquí es que hemos sido capaces de usar esas cláusulas como programas lógicos (ejecutables en un computador) para extraer tópicos y obtener resúmenes de textos especializados en español.

En la primera sección de este artículo presentamos la estrategia particular que estamos siguiendo en el proyecto. Después se presenta nuestro primer resumidor basado en la gramática de estilo y algunos ejemplos de las salidas que produce. En la siguiente sección, mostramos brevemente como funcionan esas reglas sobre un ejemplo concreto que ilustra la estrategia constructiva al resumir. Por último se concluye cómo continuará el proyecto y de las aplicaciones futuras a corto plazo.

2 La estrategia del resumidor

Investigaciones como la de Wiebe, Hirst y Horton (1996), sugieren que todo texto está asociado a un contexto lingüístico particular que determina el significado de todas sus palabras y oraciones. Según esos trabajos, el escritor de un texto se dirige a un lector con algún propósito y los lectores pueden inferir la intención subyacente y usarla para comprender el texto. En los trabajos de (Wiebe, Hirst y Horton, 1996) se concluye que el uso del lenguaje involucra mucho más que creación y comprensión de palabras aisladas. Estos autores entienden que incorporar el contexto significa expresar el matiz y el estilo en el lenguaje. La exacta escogencia de palabras, frases y estructura de las oraciones afectan el significado y el efecto preciso de una palabra. Los aspectos de estilo o enfoque del escritor son mucho más parte del mensaje que pretende el hablante que su propio significado literal.

Según los trabajos de DiMarco y Hirst (1993), un escritor usa varias construcciones sintácticas con un objetivo estilístico que ellos llaman de “alto nivel”. Con el fin de asegurar que una traducción automática retenga estos objetivos se requiere una estructura sintáctica diferente en el lenguaje destino. Para capturar esta clase de intuición lingüística, estos investigadores desarrollaron la idea de una "gramática de estilos", la cual relaciona las estructuras sintácticas de un lenguaje con un conjunto de objetivos estilísticos independientes del lenguaje. En las tareas de traducción, este objetivo puede ser determinado en el texto origen y ser usado en la generación del nuevo texto.

Por su parte, el profesor Williams de la Universidad de Chicago ha propuesto reglas de estilo para ayudar a mejorar la claridad de la escritura (Williams, 1990). Sus libros están dirigidos a los angloparlantes pero las reglas se pueden expresar en lógicas y nuestros experimentos, incluyendo los que se presentan en este artículo, parecen sugerir que se adaptan bien al español. Las recomendaciones de Williams consideran cuidadosamente las necesidades del lector.

Las sugerencias de Williams dan fundamento a ciertas reglas para la escritura de documentos. A estas reglas las llamamos “reglas de estilo” y podemos identificar los tópicos de cada oración en los textos escritos con algún apego a esas reglas. Nuestra intención es usar los tópicos como información básica para extraer descriptores significativos de los párrafos en un texto y posteriormente un resumen.

2.1 Las reglas de estilo de Williams

En Williams (1990) se presentan elementos indispensables para obtener un estilo de escritura legible: claridad, cohesión, coherencia, énfasis, elegancia, concisión y longitud. Los primeros tres son los más útiles y mejor planteados por Williams y los que usaremos en nuestro experimento. Según Williams son estos precisamente los elementos básicos de una escritura legible, que encuentra un lenguaje útil en la comunicación.

Las recomendaciones de estilo de Willians comienzan en el ámbito de la oración, luego se refieren al párrafo y por último a textos completos. A cada oración de un texto se le asocia un tópico y estos tópicos son usados para generar una secuencia coherente de oraciones que constituye un párrafo.

Williams cree que los lectores encuentran a las oraciones fáciles de leer y entender cuando su forma de razonar sigue la lógica de la oración: los sujetos de la oración deberían ser los actores, y los verbos de las oraciones deberían ser las acciones cruciales. Esto se denomina claridad, pues nos permite identificar de manera precisa los actores y acciones del relato. Para cumplir con este principio es suficiente usar una forma clara y transparente al presentar los verbos y los sujetos (Williams, 1990).

Se entiende por cohesión a “la manera como las diversas oraciones que conforman un texto

escrito permanecen unidas bajo un mismo contexto o discurso" (Williams, 1990). De esta manera, el comienzo de una oración debería retomar el pasado y conectar al lector con las ideas que se habían mencionado antes. El final de la oración debería inducir y es el lugar para colocar nuevas ideas y nueva información.

La propuesta de Williams continúa a nivel del párrafo. Las oraciones que constituyen un párrafo deberían tener tópicos consistentes y coherentes entre sí. Nuevos tópicos y nuevos temas deberían encontrarse al final de las oraciones introductorias del párrafo. Los lectores encontrarán a un párrafo como coherente si este tiene solo una oración que exprese el resumen, la cual casi siempre se encuentra o al final del párrafo o como la última de las oraciones introductorias del párrafo (esta es la clave para la coherencia).

En este trabajo, hemos construido programas lógicos para procesar textos que sigan estas recomendaciones. Si un texto atiende a esas reglas de estilo, el procesamiento lingüístico se vuelve más tratable, pues contamos con otros elementos para procesar automáticamente el contenido de los textos.

2.2 Estructura del resumidor

El resumidor implementado tiene seis componentes básicos: tokenizador, gramática, claridad, cohesión/coherencia, tópico común y salida. En la tabla 1 mostramos un esquema de estos componentes con sus entradas y salida. La programación de este resumidor fue realizada en PROLOG, un lenguaje de programación de alto nivel que permite al programador concentrarse en la lógica de su problema, antes que en los medios de ejecución particulares del computador.

Entrada	Componente	Salida
Texto	Tokenizador	Texto segmentado
Texto segmentado	Gramática	Texto etiquetado
Texto etiquetado	Claridad	Lista tópicos [1]
Lista tópicos [1]	Cohesión/ Coherencia	Lista tópicos [2]
Lista tópicos [2] + factor resumen	Tópico común	Lista tópicos [3]
Lista tópicos [3] + Texto segmentado	Salida	Texto Resumen

Tabla 1: La estructura del resumidor simbólico.

El tokenizador textual segmenta el texto en unidades de procesamiento. Es decir, identifica el párrafo como unidad de extracción y separa sus oraciones y palabras.

El siguiente componente implementa un *parser* para el idioma español en base a una gramática simplificada. La salida que se obtiene al aplicar la gramática es denominado "texto etiquetado" porque cada una de sus oraciones contiene marcas para identificar el sujeto, verbo y complemento. Este resultado se obtiene a partir de una revisión superficial de la gramática de cada oración (superficial porque apenas separa esos componentes). En esa revisión, el verbo principal es identificado utilizando un diccionario de verbos para el español, que hemos circunscrito a ciertos dominios procurando mejor cobertura para los textos procesados.

Es importante aclarar que nuestra intención no es realizar un análisis exhaustivo de todos los constituyentes gramaticales de la oración. Nuestro objetivo es identificar solamente aquellos constituyentes que nos permitirán aplicar las reglas lógicas de los siguientes niveles.

El componente "claridad" se refiere a la claridad oracional de Williams (1990). En esta fase se identifica el tópico de cada oración por separado y se filtran los conectores o marcadores discursivo[1]. Williams, define tópico como "el sujeto psicológico de la oración". Es decir, al parecer, el tópico es la parte de la oración que lleva la carga lógica del discurso, tanto oral como escrito. Progresivamente a lo largo del texto "estas ideas ´topicalizadas´ proporcionan avisos temáticos que enfocan la atención del lector hacia un conjunto bien definido y limitado de ideas conectadas".

Nosotros hemos preferido considerar en los tópicos los conceptos emitidos o involucrados en cada una de las proposiciones que posee un argumento. Desde el punto de vista sintáctico, el tópico generalmente es expresado en una frase nominal, que el resto de la oración explica o caracteriza. Considerando esto escogemos como tópico las frases nominales contenidas en

[1] Para esto se dispone de un diccionario de conectores o marcadores discursivos que fueron obtenidos en base a sugerencias de Williams (1990) y a partir del corpus de prueba.

el sujeto y en el complemento de la oración dependiente del tipo de verbo[2].

A nivel de la semántica del discurso encontramos el siguiente componente que usa las reglas de estilo definidas a nivel de párrafos -reglas de cohesión y coherencia de Williams (1990)-. Se consideraron los elementos superficiales para eliminar la ambigüedad como la repetición parcial o total de tópicos. También se tomaron en cuenta los elementos para compactar la superficie como las formas pronominales (anáforas) (Beaugrande y Dressler, 1997). El procesamiento de estos elementos esta basado en la lista de tópicos de claridad (lista tópicos [1]), para obtener una lista de tópico de cohesión (lista tópicos [2]) de menor cardinalidad.

Hasta este punto, los componentes del resumidor representan una teoría de apoyo para caracterizar a los tópicos relevantes. Es decir, son una axiomatización que nos dice que ciertas frases de ciertos discursos son tópicos relevantes o adecuados de esos discursos (Lista tópicos [2]). Hasta aquí intervienen las reglas de estilo de Williams.

Adicionalmente, hemos podido extender esta teoría de apoyo para incluir un mecanismo de ponderación de tópicos que nos permitirá obtener un resumen de cada párrafo. Esto consiste básicamente en aumentar la "prioridad" de cada tópico de acuerdo a la ocurrencia de los elementos superficiales de la cohesión textual (Beaugrande y Dressler, 1997) resueltos en el componente anterior. Luego se considera el factor de resumen (el cual varía de 0 a 2, de menor a mayor capacidad de síntesis respectivamente) para escoger los tópicos que representarán a cada párrafo (Lista tópicos [3]).

Para finalizar, el último componente genera un resumen por párrafo del texto original. Esto consiste en mostrar como resumen al conjunto de aquellas oraciones del texto que contiene a los tópicos de la lista de tópicos [3]. Aquellas oraciones que no contengan ningún tópico de la lista de tópicos final es eliminada de la salida. Cada oración resumen tiene una marca o etiqueta en el tópico para diferenciarlo del resto de la oración.

La estrategia descrita muestra como nuestro resumidor es capaz de extraer del texto original las oraciones que contiene a los tópicos del párrafo. Esto también se observa gráficamente en la figura 1.

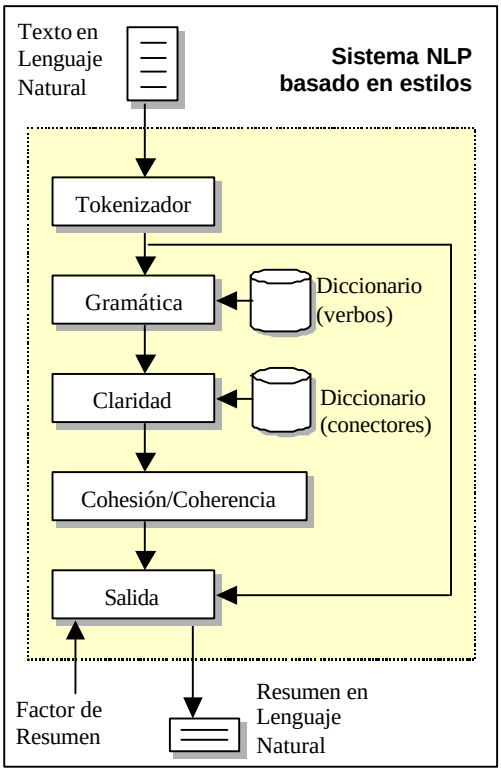

Figura 1. Un diagrama del resumidor basado en una "gramática de estilos".

3 Ejemplo del funcionamiento del resumidor

Mostramos a continuación, en (1), un párrafo de uno de lo artículos analizados y seguidamente, en (2), el resumen producido por el programa:

(1) El deterioro de la actividad agrícola se tradujo en una menor participación del sector en la actividad económica, debido principalmente al surgimiento de la industria petrolera. En consecuencia, dentro del mencionado sector, la declinación de la actividad cacaotera fue incontenible. El Estado crea el Fondo Nacional del Café y del Cacao (FNCC) en 1959, con el objetivo de promover el cultivo y, a la vez, controlar la comercialización de ambos rubros. Esto formó parte de las políticas proteccionistas gestadas en esa época. En 1975 se

[2] Se toman en cuenta tres tipos de verbos: auxiliares, predicativos e impersonales (con pronombre impersonal "se").

reestructura el Fondo Nacional del Café y del Cacao, dividiéndose en dos organismos autónomos e independientes: el Fondo Nacional del Café (FONCAFE) y el Fondo Nacional del Cacao (FONCACAO). De esta forma, se inicia el monopolio de la compra, distribución y exportación del cacao, ejercido por el Estado a través de FONCACAO como empresa comercializadora.

(2) En consecuencia, dentro del mencionado sector, **la declinación de la actividad cacaotera** fue incontenible. **El Estado** crea **el Fondo Nacional del Café y del Cacao (FNCC)** en 1959, con el objetivo de promover el cultivo y, a la vez, controlar la comercialización de ambos rubros. Esto formó **parte de las políticas proteccionistas gestadas en esa época**. De esta forma, se inicia **el monopolio de la compra**, distribución y exportación del cacao, ejercido por el Estado a través de FONCACAO como empresa comercializadora.

El texto anterior, es reducido de 140 palabras a las 83 mostradas en el ejemplo (2). Observe que el párrafo conserva algo de sentido y contiene material relevante, aún cuando el resumidor no posee conocimiento experto sobre el dominio.

Las reglas de estilo de Williams nos permiten explicar el resultado anterior. En primer lugar, las reglas de coherencia indican que las ideas que se colocan al inicio de un párrafo deben ser información conocida por el lector o referida anteriormente en el texto. Esta información es lo que se denomina el arranque de un párrafo coherente (Williams, 1990). Las características del arranque nos lleva a considerarlo como información menos relevante para el resumen del párrafo. Por esta razón, la primera oración del texto (1) es eliminada del texto (2).

Luego del arranque del párrafo se tiene una serie de oraciones que introducen cadenas temáticas en relación al tópico arranque. Esto es lo que se denomina la discusión de un párrafo coherente (Williams, 1990). Los tópicos (destacados en negritas) contenidos en la discusión se obtienen aplicando las reglas de claridad y cohesión. En base a ellos y a las reglas de "tópico común" se eliminó la quinta oración del texto original. Como se puede observar, el tópico de esta oración es mencionado anteriormente (Fondo Nacional del Café y del Cacao), agregando información nueva. En particular, la repetición parcial o total de tópicos nos permite reducir el contenido del resumen escogiendo solamente su primera ocurrencia.

Por otra parte, nuestro resumidor puede recibir como parámetro un mayor nivel de síntesis (factor de resumen igual a 2), permitiendo obtener sobre el mismo ejemplo la siguiente salida:

(3) El Estado crea **el Fondo Nacional del Café y del Cacao (FNCC)** en 1959, con el objetivo de promover el cultivo y, a la vez, controlar la comercialización de ambos rubros.

No pretendemos reportar mayor efectividad con el programa actual[3]. Sin embargo, observamos que la oración del ejemplo (3) es apropiada para transmitir el sentido de ese párrafo, sin entrar en detalles. Nuestro reto ahora es afinar el resumen con criterios de relevancia más precisos, sin desmejorar los logros que ya se obtienen. Creemos que esto es posible, pues la representación se presta a la elaboración y a la integración con otros mecanismos.

4 Evaluación del resumidor basado en el estilo

Un resumidor, en la intuición de cualquier lingüista, debería tener en cuenta el concepto de relevancia o de un substituto apropiado, respecto a algún contexto o solicitud de información. Una noción de la relevancia ha sido propuesta por Cooper (1971) y la denomina "relevancia lógica". La relevancia está definida en términos de la consecuencia lógica. Un documento es relevante a una necesidad de información, si y solamente si, contiene por lo menos una sentencia que sea relevante a esa necesidad.

Los investigadores de esta área han destacado la importancia de producir metodologías de evaluación de sistemas NLP,

[3] Página con interfaz web para el resumidor http://cesimo.ing.ula.ve/INVESTIGACION/PROYECTOS/GIL/esumidor.html

porque pueden resultar muy costosos y complejos (King, 1996). King dice que los resultados varían enormemente en función del propósito, del alcance, y de la naturaleza de los objetos que están siendo evaluados.

Tomando esto en consideración, debemos acotar la necesidad de información y el tipo de resumen que pretendemos producir (Acero et al., 2001) (Hahn y Mani, 2000). En este resumidor, el alcance está limitado a procesar un único documento. Además se trata de un resumen *indicativo*, pues el objetivo es anticipar al lector sobre el contenido del texto y ayudarle a decidir sobre la relevancia del mismo. Por último, el enfoque es *genérico*, pues recoge los temas principales del texto, sin sesgarlos a un grupo de usuarios en particular.

Según Maña, Buenaga y Gómez (1998), las técnicas empleadas en la generación de resúmenes pueden ser estadísticas o simbólicas, y utilizadas en aplicaciones generales o específicas, respectivamente. Los sistemas estadísticos, independientes del dominio, generan resúmenes inconsistentes e incompletos. Sin embargo, se han mejorado estos resultados incorporando reglas semánticas y discursivas, agregándoles también cierta dependencia al contexto y tipología textual específica.

En nuestro caso se trata de un resumidor que usa técnicas simbólicas basadas en reglas discursivas y de estilo. Para el experimento se seleccionó un dominio de conocimiento particular y se escogió un conjunto de textos especializados de una revista académica local[4]. Esto nos ha permitido contar con algunos expertos en el dominio del contenido del texto para comparar la relevancia obtenida por el sistema.

En esta versión preliminar del resumidor no se ha incluido ningún modelo explícito del área de conocimiento en la cual están suscritos los textos, es decir, no se dispone de una base de conocimiento para el dominio, ni menos de reglas que integren este conocimiento al procesamiento gramatical y estilístico. Sin embargo, el diccionario de verbos ha sido construido específicamente para el dominio de conocimiento sobre el cual operará el resumidor. Esto se puede considerar un primer paso en la adecuación y circunscripción del sistema a un dominio particular.

Para la evaluación usamos un corpus textual que presenta parcialmente los elementos de estilos de Williams (1990). En base a este corpus textual se construyó un diccionario de verbos (anotados por categorías verbales), para cuya construcción se empleó la herramienta TACT (Textual Analysis Computing Tools)[5].

Los evaluadores que revisaron los resúmenes informaron sobre la pertinencia de los mismos para tener una idea preliminar del contenido completo del documento. Sin embargo, algunos destacaron la existencia de oraciones aisladas que no parecían coherentes. En otros casos existe información temporal que no es tomada en cuenta dentro de la secuencia de las oraciones. Además algunos textos resúmenes presentaron referencias anafóricas y elipsis que no permiten la total comprensión del contenido textual. Los evaluadores destacaron como positivo la manera en que se destacaron los tópicos dentro del texto resumen.

5 Limitaciones y escalabilidad

Nuestro programa resumidor está dirigido a a cierto tipo de textos especializados. Estos textos tienen generalmente rasgos morfosintácticos que involucran las referencias verbales, sustantivos, y adjetivos descritas en Amtz y Picht (1995). Generalmente, los textos con estas características también se ajustan a las sugerencias de estilo de Williams. Sin embargo, los textos no especializados pueden aplicar si consideran estas reglas de estilo. Más aún, el resumidor puede procesar un texto que no se ajuste al estilo Williams. Nuestro objetivo de robustez es que, en estos casos, el programa no colapse y genere una salida, aún si no es muy útil como resumen.

Otra limitación está relacionada a los diccionarios empleados en el resumidor. Se usa un diccionario de verbos en español para el componente gramatical y un diccionario de marcadores discursivos para las reglas de claridad. Ambos diccionarios fueron generados

[4] Revista "Agroalimentaria" del (CIAAL) Centro de Investigaciones Agroalimentarias, Facultad de Ciencias Económicas y Sociales de la Universidad de Los Andes. Mérida – Venezuela.

http://www.saber.ula.ve/ciaal/agroalimentaria/

[5] TACT es un sistema de recuperación y análisis textual sobre bases de datos textuales en idiomas europeos. Este sistema se comenzó a desarrollar como una iniciativa de cooperación hacia las Humanidades por parte de IBM y la Universidad de Toronto durante los años 1986-89.

http://www.chass.utoronto.ca/cch/tact.html

a partir del corpus de prueba. Por tanto, el sistema descarta aquellas oraciones cuyo verbos no están en el diccionario verbal. Además, las oraciones que contengan marcadores discursivos que no estén en el diccionario serán incluidos como parte del tópico.

Sería muy importante considerar reglas de estilo a nivel de texto completo o secciones de texto, pues todas las reglas de estilos aplicadas son localizadas a nivel de párrafo. La claridad y la cohesión aplican a nivel de oración y párrafo. La coherencia va mas allá y considera las relaciones entre los párrafos, pero estas relaciones son generalmente conceptuales, y nuestro sistema debe representar conocimiento de cada dominio si queremos que aplique para estas tareas.

5.1 Hacia resúmenes constructivos.

Uno de nuestros objetivos pendientes es un resumidor constructivo más general. Es decir un resumidor capaz de obtener como resultado oraciones o frases nuevas que no estén literalmente en el texto, pero que representen un resumen del texto con frases u oraciones correctas en español.

Según Maña, Buenaga y Gómez (1998) las técnicas de resumen que son constructivas suelen estar circunscritas a dominios muy concretos. La idea que tenemos es explotar la semántica de los verbos, la cual tiene mayor independencia del dominio de conocimiento, con el fin de definir el significado de una oración. Esto quiere decir que además de determinar el tópico de las oraciones según el tipo de verbo, como se menciono en el componente de claridad, se puede usar el significado del verbo. Una implementación de esta idea consistiría en usar a un "tesauro de verbos y sus representaciones semánticas" y según esos significados escoger como tópico a cierta parte de la oración. De esta manera, contando con la semántica de los verbos y sus tópicos podríamos relacionar los tópicos de las oraciones con ciertos conectores.

Para ilustrar esta idea, considere el siguiente texto de una noticia internacional, tomada de "The Wall Street Journal Americas" (15-02-2001).

El texto del discurso es el siguiente: "Un informe de un comité científico de la unión Europea reveló que las ovejas y las cabras pueden contraer, teóricamente, el mal de las vacas locas. Pero que hasta ahora esto solo ha ocurrido en experimentos de laboratorio".

Las reglas de extracción son las siguientes:

R1- T es un tópico del discurso D si en el discurso D, un Agente *revela* T

R2- T es un tópico del discurso D si en el discurso D, un Agente *revela* T **y** T contiene la información que Agente2 *puede* hacer T.

R3- T es un tópico del discurso D si en el discurso D, un Agente *revela* T **y** T contiene la información que Agente2 *puede contraer* Algo y T = Algo **en** Agente2.

R4- Un tópico T es el común más específico en D si es un tópico del discurso D **y no** existe otro tópico T de D tal que T mas general que T.

R5- T es más general que T si T es más breve que T.

Con estas reglas el tópico T resultante es: "mal de las vacas locas **en** ovejas y cabras".

Un conjunto de reglas como las anteriores fueron incorporadas al resumidor para producir la frase resumen al final del párrafo anterior. Observen que, si bien estas reglas fueron inspiradas por ese texto en particular (y permiten resolverlo), las reglas son generales: pueden aplicar a otras oraciones en donde se cumpla la peculiar relación de un tópico "en" otro que modelan esas reglas.

5.2 Paradigmas en las relaciones entre tópicos

Esa última observación nos ha llevado a considerar una estrategia para ampliar la cobertura del resumidor y permitir la generación "constructiva" de resúmenes. La estrategia consiste en definir "paradigmas" en las posibles relaciones entre tópicos en una misma oración, como el caso del último ejemplo, y relaciones entre tópicos en oraciones distintas (en un mismo párrafo, para comenzar).

Observen que en cuanto a esos paradigmas inter-oracionales existe uno especial que ya hemos incorporado, en una primera forma, al resumidor actual. Es el paradigma de "Tópico1 igual a Tópico2" que nos permite decidir si un tópico dado se repite a lo largo de un párrafo. Como podrán apreciar, esa igualdad no es, salvo en casos triviales, una igualdad sintáctica.

La comparación apela a la semántica puesto que, quizás en procura de elegancia, los escritores rara vez repiten exactamente la misma frase tópico.

6 Conclusiones

El problema que abordamos en este artículo es el de interpretar textos en español y extraerles sus tópicos. Hemos argumentado nuestra creencia de que es posible explotar criterios lógicos para asociar tópicos adecuados y relevantes a textos en español y hemos ilustrado el argumento con las salidas de un programa lógico Este programa, codificado en PROLOG, instrumenta el proceso de análisis del lenguaje natural que hemos venido explicando. Si bien este código se concentra en el análisis de oraciones y párrafos, con los criterios de claridad, cohesión y coherencia de Williams, se le puede considerar una versión preliminar de un resumidor.

En esta propuesta se implementaron: (1) reglas lógicas inspiradas en los estilos de Williams, (2) criterios de tópicos relevantes expresados también como reglas y (3) una representación del dominio de conocimiento, la cual, por los momentos, es un diccionario de verbos elemental, pero que puede extenderse con información semántica y pragmática verbal y también puede incorporarse ontologías y diccionarios terminológicos del dominio de conocimiento.

Nuestra principal hipótesis de trabajo establece que, si se aplica una gramática adecuada a ciertos estilos, entonces se pueden obtener sistemáticamente los tópicos adecuados de un documento escrito sobre la base de esos estilos. Se esperaría que los tópicos derivados con esta estrategia sean próximos a los descriptores obtenidos por los expertos en el dominio. Resta todavía mucho trabajo para aproximar esos descriptores a los producidos por expertos humanos. Sin embargo, es motivador reportar que el formalismo para representación del conocimiento lingüístico que estamos empleando (Lógica) ha sido tolerante a la elaboración de reglas de análisis basadas en criterios poco ortodoxos en el procesamiento lingüístico (las reglas de estilo).

Bibliografía

Acero, I., Alcojor, M., Díaz A. y Gómez, J.M. 2001. Generación automática de resúmenes personalizados. Procesamiento del Lenguaje Natural, 27.

Amtz, R. y Picht, H. 1995. Introducción a la Terminología. Madrid: Fundación Germán Sánchez Ruipérez. Traducido por Irazazábal A et al.

Beaugrande, R.A. y Dressler, W. U. 1997. Introducción a la Lingüística del texto. 1ª edición en español. Editorial Ariel, S.A. Barcelona España. ISBN: 84-344-8215-0.

Cooper, W.S. 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7, 19-37.

DiMarco, C. y Hirst, G. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*. 19, 3 Septiembre 451-499.

Hahn, U. y Mani, I. 2000. The Challenges of Automatic Summarization. *IEEE Computer*, Noviembre, 33(11):29-35.

King, M. 1996. Evaluating Natural Language Processing Systems. *Communications of the ACM*. Enero, Vol. 39, No. 1.

Maña, M., Buenaga, M. y Gómez J.M. 1998. Diseño y evaluación de un generador de resúmenes de texto con modelado de usuario en un entorno de recuperación de información. En *XIV Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN'98)*, 23-25 septiembre, Alicante (España). Publicado en *Procesamiento del Lenguaje Natural*, nº 23, septiembre, 32-39.

Pereira, F.C.N. y Warren, D.H.D. 1986. Definite clause grammars for language analysis-a survey of the formalism and a comparison with augmented transition networks. *Artificial Inteligence*. 13:231-278. pag. 101-138.

Wiebe, J., Hirst, G. y Horton, D. 1996. Language use in Context. *Communications of the ACM*, Enero 1996, Vol. 39, No. 1.

Williams, J. 1990. *Style: Toward Clarity and Grace*. The University of Chicago Press. Chicago and London.