CENTRO DE LINGÜÍSTICA APLICADA
MINISTERIO DE CIENCIA TECNOLOGÍA, Y MEDIO AMBIENTE
SANTIAGO DE CUBA

ACTAS - II

IX SIMPOSIO INTERNACIONAL DE COMUNICACIÓN SOCIAL
SANTIAGO DE CUBA, 24 - 28 DE ENERO DE 2005

Edición y Composición:
Celia Álvarez Moreno
Jorge Pérez Bolaños
Laritza Hernández Rojas
Leonel Ruiz Miyares
Yilian Cortés Gutiérrez

ISBN: 959-7174-05-7

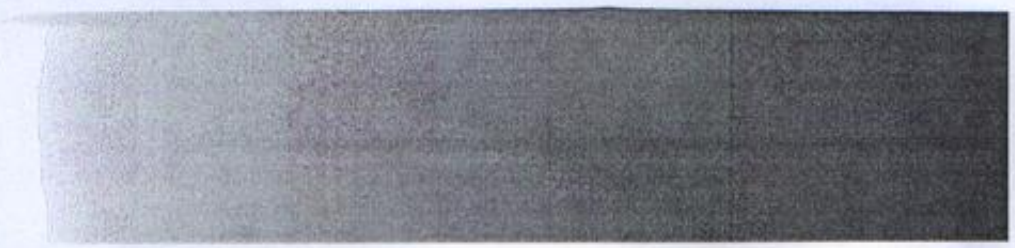
COAUSPICIADORES

- Universidad de
- Universidad Pec
- Dirección Provir
- Centro Cultural
- Teatro Heredia,
- Hotel Meliá – Sa
- Universidad de
- Knowledge Wel
- Universidad del
- Instituto de Ling
- Universidad de
- Agencia de Viaj

COMITÉ ORGANIZAD

Presidenta de Honor:
Dra. Rosa Elena Sime
Ministra de Ciencia, T
República de Cuba

Pedro Beatón Soler
Eloína Miyares Bermú
Vitelio Ruiz Hernánde:
Leonel Ruiz Miyares
Zaida Valdés Estrada
Ena Elsa Velázquez C
Anton Nijholt
Nancy Cristina Álamo
Celia Pérez Marqués
Celia Álvarez Moreno
Mercedes Cathcart Ro
Ercilia Estrada Estrad
Martha Cordiés Jacks
Miladys Diodene Adan
Dieter Fensel
Iñaki Alegria Loinaz
Arantza Díaz de Ilarraz
Xabier Artola Zubillag
Xabier Arregi Iparragir
Asa Abelin
Jens Allwood
Nicoletta Calzolari
Daniela Ratti
Lucia Marconi
Claudia Rolando
Paola Cutugno
Raúl Ávila
Michael Zock
Sylviane Cardey
Peter Greenfield
Félix Rodríguez
Ruslan Mitkov



ÍNDICE

MARÍA MARILÚ PARRA
JACINTO DÁVILA
ULA, Mérida, Venezuela
mmarilu@ula.ve

Un modelo computacional para la generación de resúmenes automáticos de artículos científicos en español

Resumen[1]

Los propósitos generales del lenguaje escrito[2] son básicamente los mismos independientemente de la lengua en que se genere. ***Así como la necesidad*** de ser comprendido por ***otros es universal***, las sociedades necesitan transmitir su herencia de ideas y conocimientos a través del lenguaje. En este trabajo definimos el conocimiento como un flujo mixto de experiencia, valores, información contextualizada y visión experta que provee un marco de referencia para evaluar ***e incorporar nuevas experiencias e información***. Desde este punto de vista, el conocimiento es producto de un proceso dinámico y como tal, se fundamenta en gran medida de la transmisión de la información. En la actualidad la gran cantidad de información que se genera en forma de texto, y que está disponible gratuitamente en ***Internet*** se incrementa día a día. Esto hace que se encaminen esfuerzos avocados a la búsqueda de soluciones en el procesamiento del lenguaje natural, lo que ha contribuido a ***impulsar la investigación y el desarrollo de técnicas y aplicaciones que combinan tecnología con conocimiento lingüístico***, monolingüe y multilingüe, dando lugar a la llamada ingeniería lingüística. El objetivo principal de esta área de estudio es la aplicación del conocimiento de la lengua al desarrollo de sistemas informáticos capaces de reconocer, comprender, interpretar y generar el lenguaje humano en todas sus formas. Estas técnicas permiten "entender" el ***texto*** o el habla en lenguaje humano y desarrollar tareas que requieren de tal comprensión. El dictamen principal que rige nuestra conceptualización de lo que hemos definido como proceso de comprensión textual, se fundamenta en que los textos deben abordarse no sólo como ***un conjunto de oraciones***, sino como un todo con sentido completo.

La formalización del proceso de comprensión está orientada al ***reconocimiento de la superestructura***, específicamente al reconocimiento de las secciones retóricas de carácter persuasivo para obtener como resultado un resumen automático del texto. Para ello fue necesario ***la delimitación del contexto***. Dicha limitación estuvo dirigida a la identificación de la estructura de Artículos de Investigación Científica (AIC), con el objeto de ***enfocarnos en aspectos específicos de este tipo de literatura***. El modelo que explicita la superestructura de los AIC es representado a través ***del modelo IMRD (introducción, métodos, resultados, discusión)*** y el modelo Swales [Swales, 1990] para las introducciones. ***Con la aplicación de estos modelos teóricos y con la ayuda de nuestro corpus de estudio que sirvió de banco de pruebas*** fue posible sistematizar el proceso que siguen los humanos en la producción y consumo de textos especializados. A través de ***la sistematización*** y haciendo uso de lenguajes formales fue posible formalizar el proceso de reconocimiento de superestructura de los AIC. Sin embargo, aún se presentan ***ciertas limitaciones*** en ***estos intentos por teorizar las estructuras textuales***. Al igual que las técnicas usadas en el Procesamiento del Lenguaje Natural (PLN), estos esfuerzos están encaminados a facilitar y mejorar la comunicación humana, por lo que se hace necesario la interdisciplinariedad para alcanzar este objetivo. Sin embargo, los avances en este campo están aún muy lejos de lograr ***resultados*** óptimos o equivalentes al rendimiento humano.

Al final de este trabajo planteamos una propuesta para extender las bondades de este programa, la cual se enfoca en la ***formalización de las macrorreglas*** semánticas para generar un resumen automático producto de la inferencia de las proposiciones explícitas en el texto.

El problema investigado

El texto **Resumen** que Ud. acaba de leer fue generado automáticamente a partir de un extenso documento (más de 120 páginas) dispuesto con una estructura documental similar a la de un artículo científico. Se trata de un ejercicio extremo de la lingüística computacional del texto que hemos conducido con la intención de validar un modelo de la superestructura de artículos científicos en Español [Parra, 2004].

La proliferación de información en formato digital motiva el trabajo. A menudo, para poder hacer uso eficiente de la información del texto, se requiere que la misma esté puesta en una cierta clase de formato estructurado. Extraer la información requerida de entre grandes volúmenes de documentos es, generalmente, un proceso manual costoso; por lo tanto, el texto en formato digital crea una necesidad de métodos de procesamiento para extraer la información automáticamente [Jacobs, 2000].

En este trabajo ahondamos en el problema de la formalización del proceso de comprensión para la generación de resúmenes, ajustado a un contexto específico de textos escritos con cierto propósito y estilo y con la intención de construir una plataforma teórica, computacionalmente factible, para el análisis de textos.

Objetivo de la investigación

El propósito general de este proyecto fue crear un modelo de la estructura de un texto que podamos usar luego para guiar un sistema de cómputo que aproveche esa estructura para analizar y sintetizar (resumir) un texto. Ese modelo puede entenderse como un refinamiento o una formalización de la teoría de estilos para escribir textos claros y comprensibles propuesta por J. Williams [Williams, 1990]. Ese trabajo de formalización comenzó con un

---

[1] Este resumen fue generado automáticamente (nivel 0) de una versión del documento en formato HTML.
[2] Las frases con este estilo de edición (negrita e itálica) son los tópicos identificados por el resumidor durante la generación del resumen automático.