UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERÍA
CONSEJO DE ESTUDIOS DE POSTGRADO
POSTGRADO EN MODELADO Y SIMULACIÓN

UN MODELO COMPUTACIONAL PARA LA GENERACIÓN DE RESÚMENES AUTOMÁTICOS DE ARTÍCULOS CIENTÍFICOS EN ESPAÑOL

Trabajo de grado presentado como requisito parcial para
optar por el grado de MAGISTER SCIENTIAE en
Modelado y Simulación de Sistemas

Marilú Parra
Tutores: Jacinto Dávila
Françoise Meyer
Melva Márquez

Mérida – Venezuela
Julio 2004

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

[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.

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.

Sócrates - Pero al menos me concederás que todo discurso debe, como un ser vivo, tener un cuerpo que le sea propio, con pies y cabeza, medio y extremidades debidamente proporcionadas entre sí y con el conjunto.

Fedro - Evidentemente.

Platón

1

Referencias

[Adam, 1992] Adam, J. M. Les Textes: Types et Prototypes. En Propuesta de Utilización de Modelos Textuales Definidos Culturalmente para La Enseñanza de la Lectura en Inglés y Español. Trujillo Sáez, Fernando Ed. Facultad de Educación y Humanidades de Ceuta, Departamento de Didáctica de la Lengua y la Literatura. Universidad de Granada. 2002.

[Aguado, 2001a]. Aguado, Guadalupe. 2001. "Creación terminológica y desarrollo científico en informática". En: Aportaciones de la ingeniería a la lengua Española. Instituto de Ingeniería. España. 2001.

[Albadejo, 1987] Albadejo Mayordomo, Tomás. Componente Pragmático, Componente de Representación y Modelo Lingüístico Textual. En Lingüística del texto. Bernández, Enrique. Ed. Arco Libros, Madrid, 1987.

[André-Haudricourt 1992], André, Haudricourt. *Estructuralismo y Linguística*. Nuevisión, Argentina. 1992.

[Astorga, 2003] Astorga, Luis. *Simulación Lógica de Relatos Narrativos*. Tesis de Grado de Maestría, Posgrado en Modelado y Simulación de Sistemas, Universidad de Los Andes. 2003.

[Atkinson, 1993] En *"The Schematic Structure of Computer Science Research Articles"* Posteguillo, Santiago. English for Specific Purpose. Volume 18. 1999.

[Bajtín, 1985]. Bajtin, M.M. El Problema de los Géneros Discursivos. En Estética de la Creación Verbal. Siglo XX, Madrid. 1985.

[Bernández, 1982] Bernández, Enrique. *Modelos de Lingüística Textual*. En *Introducción a la Linguístia del Texto*. Espasa-Calpe, Madrid, 1982.

[Bustos, 1996]. Bustos Gisbert, José. *La Construcción de Textos en Español*. Universidad de Salamanca. España, 1996.

[Cabré, 1997] Cabré, María T. *Lenguaje General y Lenguajes de Especialidad: Palabras y Términos*. Antartida/Empúries, Barcelona. 1997.

[Cabré, 2001] Cabré, María T. *La Terminología Científico Técnica: Reconocimiento Análisis y Extracción de Información Formal y Semántica*. Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. Barcelona. 2001

[Callaway-Lester, 2002] Callaway, Charles y Lester, James. *Narrative Prose Generation*. Artificial Intelligence. Department of Computer Sciencie. 2002.

[Chapa, 1999] Chapa, Sergio. *Introducción a la lógica matemática*. Centro de Investigación y de Estudios Avanzados del IPN. 1989.

[Colle, 2000] Colle, Raymond. *Análisis de Contenido*. Curso a Distancia. Universidad de Chile. 2000. disponible en: http://www.puc.cl/curso_dist/conocer/analcon/introd.html.

[Contreras, 2002] Contreras, H. *Una Técnica para la Extracción Automática de Resúmenes Basada en una Gramática de Estilo*. Tesis de Grado de Maestría, Universidad de Los Andes. Mérida-Venezuela. 2002.

[Contreras-Dávila, 2001] Contreras, H. y Dávila, J. *Procesamiento del Lenguaje Natural Basado en una "Gramática de Estilos" para el Idioma Español*. Memorias de la Conferencia Latinoamericana de Informática, CLEI. Mérida-Venezuela. 2001.

[Covington, 1993] Covington, Michael A. *Natural Language Processing for Prolog Programmers*. Prentice Hall, Englewood Cliffs. New Jersey 07632. 1994.

[Dávila, 2001]. Dávila, Jacinto. *Modelos: Una Visión Integradora*. Reporte del Centro de Simulación y Modelos (CESIMO). Universidad de Los Andes. 2001.

[Dávila, 2003]. Dávila, Jacinto. *Curso de Lógica Matemática*. Maestría en Modelado y Simulación, Universidad de los Andes. Mérida-Venezuela. 2003. Disponible en http://www.saber.ula.ve.

[Dávila-Contreras, 2002]. Dávila, J. y Contreras, H. *Una Gramática de Estilos para Resumir Textos en Español*. XVIII Congreso de la SEPLN, Septiembre 2002. Valladolid, España. 2002.

[Dávila-et-al, 2002]. Dávila, J.; Astorga, L.; Márquez, M., Contreras, H., Myerston, J. y Parra, M.. "*Introducción a la Lingüística Computacional con una Perspectiva Interdisciplinaria*". Terminómetro. Ed. Unión Latina. Número. 6. Paris. 2002.

[de Beaugrande, 1987] Beaugrande, Robert. *Teoría Lingüística y Metateoría para una Ciencia del Texto*. En *Lingüística del texto*. Bernández, Enrique. Arco Libros, Madrid, 1987.

[de Beaugrande-Dressler, 1997] De Beaugrande, R.A. y Dressler, W. U. *Introducción a la Lingüística del Texto*. 1ª edición en español. Editorial Ariel, S.A. Barcelona España. 1997.

[Drop, 1987]. Drop, W. *Planificación de Textos con Ayuda de Modelos Textuales* en Bernández, Enrique. *Lingüística del texto*. Arco Libros, Madrid, 1987.

[Ejalde, 1998]. Ejalde, Alfredo. *Discurso Literario y Discurso Académico*. Apuntes, Revista Electrónica. Perú. 1998.

[Enkvist, 1987]. Enkvist, Nils Eric. Estilística, Linguística del Texto y Composición**.** 1987.

[Fernández, 2000]. Fernández Toledo, Piedad. *Contexto Pragmático, Género y Comprensión Lectora de Resúmenes Científicos en Inglés*. Filología Inglesa, Ciencias de la Documentación, Universidad de Murcia. 2000.

[Fredrickson-Swales,1994] Fredrickson, K. y Swales, J. *Text and Talk.* En *Profesional Contexts. International Conference "Discourse and the Professions"*. The Swedish Association of Applied Linguistics. Uppsala. 1994.

[G.E.D., 2000]. Grupo de Estructuras de Datos. *Informática Documental y Lingüística Computacional*. Departamento de Informática y Sistemas. Universidad de las Palmas de Gran Canaria. 2000.

[Gelbukh, 2002]. Gelbukh. *Tendencias Recientes en el Procesamiento de Lenguaje Natural.* En Memorias de SICOM-2002, Villahermosa, Tabasco, México. 2002.

[Gnutzmann-Oldenburg, 1991]. Gnutzmann, Claus y Oldenburg, Hermann. Contrastive Text Linguistics in LPS-Research: Theorical Considerations and some Preliminary Findings, en: Subject-oriented texts. Languages for Special Purposes and text Theory. Walter the Gruyter & Co. Berlin. Volume 16. 1991.

[Grishman, 1991] Grishman, Ralph. *Introducción a la Lingüística Computacional.* Visor Distribuciones. Madrid. 1991.

[Hahn, U. y Mani, I] Hahn, U. y Mani, I. The Challenges of Automatic Summarization. Computer IEEE. Noviembre, 2000, citado por Contreras, H. Una Técnica para la Extracción Automática de Resúmenes Basada en una Gramática de Estilo. Tesis de Grado de Maestría, Universidad de Los Andes. Mérida-Venezuela. 2002.

[Hogger, 1990]. Hogger, Christopher Jonh. Essentials of Logic Programming. Clarendon Press, Oxford, 1990.

[Hyon, 2001] Hyon, Sunny. Long-Term effects of genre-based instructions: a follow-up study of an EAP reading course, en: English for Specific Purpose. Volume 20. 2001.

[Isenberg, 1987] Isenberg, Horst. En *Introducción a la Lingüística del Texto*. Bernández, Enriquez Ed. Espasa-Calpe, Madrid. 1987.

[Lerat, 1997] Lerat, Pierre. *Las lenguas especializadas*. Ariel Lingüística. Barcelona-España. 1997.

[Meyer, 1999] Meyer, Francoise. *El Carácter Sistémico del Proceso de Comprensión: La Teoría de los Esquemas,* En: *Sistemas*. Escuela de Ingeniería de Sistemas. Universidad de Los Andes. Mérida-Venezuela. 1999.

[Moreno, 1998] Moreno Sandoval, A. Lingüística Computacional: Introducción a los Modelos Simbólicos, Estadísticos y Biológicos. Madrid. Editorial Síntesis. 1998.

[Pérez, 2002] Pérez, M. Chantal. Explotación de los Córpora Textuales Informatizados para la Creación de Bases de Datos Terminológicas Basadas en el Conocimiento. Estudios de Lingüística Española, Red Temática de Lingüística Española, Vol, 18. Universidad de Málaga. España. 2002.

[Petöfi, 1982] Petofi, en En *Introducción a la Lingüística del Texto*. Bernández, Enriquez Ed. Espasa-Calpe, Madrid. 1987.

[Picas, 2002] Picas Vidal, Josep M. *Gestión del Conocimiento*. VIII Congreso Nacional de Informática Médica. Madrid-España. 2000.

[Posteguillo, 1999] Posteguillo, Santiago. "The Schematic Structure of Computer Science Research Articles" en: English for Specific Purpose. Volume 18. 1999.

[Salager-Meyer-et-al, 2003] Salayer-Meyer, F.; Alcaraz Ariza, M.; Zambrano, N. The Scimitar, the Dagger and the Glove: Intercultural Differences in the Rhetoric of Criticism in Spanish, French and English Medical Discourse (1930-1995) en: English for Specific Purposes. The American University. 2003.

[Sidorov, 2001] Sidorov, Grigory. Problemas Actuales en Lingüística Computacional. Revista Digital Universitaria. Universidad Nacional Autónoma de México, Nº1, Vol.2. 2001.

[Stubbs, 1987] Stubbs, Michael. *Análisis del Discurso: Análisis Sociolingüístico del Lenguaje Natural*. Alianza. Madrid-España. 1987.

[Swales, 1990] Swales, John M. *Genre Analysis English in Academic and Research Settings*. Cambridge University Press. 1990.

[Trujillo, 2000] Trujillo Sáez, Fernando. *Propuesta de Utilización de Modelos Textuales Definidos Culturalmente para la Enseñanza de la Lectura en Inglés y Español.* Facultad de Educación y Humanidades de Ceuta, Departamento de Didáctica de la Lengua y la Literatura. Universidad de Granada. 2002.

[Van Dijk, 1983]. Dijk, Teun van. *Texto y Contexto*. Cátedra, Madrid. 1983.

[Van Dijk, 1989]. Dijk, Teun van. *La Ciencia del Texto*. Paidos, Barcelona. 1989.

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