Procesamiento del lenguaje natural basado en una “gramática de estilos” para el idioma español.

Hilda Y. Contreras Z.
Postgrado en Computación.
Universidad de Los Andes.
Mérida, Venezuela.
hyelitza@yahoo.com

Jacinto A. Dávila Q.
Centro de Investigación y Proyectos en Simulación y Modelos, Postgrado en Computación.
Universidad de Los Andes.
Mérida, Venezuela.
jacinto@ula.ve

RESUMEN

Este artículo presenta un proyecto de investigación que pretende desarrollar una herramienta para interpretar documentos en español y extraer de ellos descriptores relevantes. Los problemas de procesar el lenguaje natural y de extraer información, han sido atacados desde hace varias décadas [13], [2], [18]. Sin embargo, las investigaciones no han sido suficientes para diseñar un sistema que interprete el lenguaje natural con un rendimiento cercano al de un humano. El lenguaje natural escapa a todos los esfuerzos de tratamiento computacional, al parecer, debido a que el conocimiento lingüístico está asociado de formas sutiles y desconocidas con el conocimiento contextual que tiene el hablante [15]. En este trabajo abordaremos el problema de la interpretación del lenguaje escrito usando gramáticas de estilos y formas lógicas. La gramática de estilo se inspira en las reglas de estilo que propone *J. Williams* [16] para escribir prosa en inglés. Esta estrategia adaptada al español y la definición de un buen descriptor, tienen la finalidad de reducir la complejidad del procesamiento sintáctico/semántico tradicional; Además de incorporar el conocimiento contextual en el proceso. Validaremos la estrategia con un prototipo de un módulo de asignación de descriptores para un sistema bibliográfico virtual.
Palabras claves: Procesamiento del lenguaje natural, Lingüística computacional, Minería de datos, Recuperación de información, Inteligencia artificial.

Natural language processing based on “grammars of style” for Spanish language.

ABSTRACT

This paper presents a research project which tries to develop a system to process documents in Spanish and extract good descriptors from them. Natural languages have been the subject of numerous attempts of systematic analysis and the development of automatic, sense-extraction mechanisms [13], [2], [18]. However, all this research effort has been unable, as yet, to develop a system with human-like performance in natural language processing. It seems that natural languages escape from computational treatment due to the fact that linguistic knowledge is intertwined with some very subtle and some unknown forms of contextual knowledge easily reachable by human beings [15]. This work will concentrate efforts on written natural language processing by means of grammars of style and logic-based knowledge representation. The grammars of style that are being used are based on those proposed by Williams [16] for the English language. The style-based strategy, its adaptation to Spanish and logical definitions for good descriptors should reduce the complexity of the syntactical/semantical analysis, making the project feasible.
Keywords: Natural language processing, Computational linguistics, Data mining, Information retrieval, Artificial intelligence.

1 INTRODUCCIÓN

El proyecto pretende desarrollar una herramienta para interpretar documentos en español y extraer de ellos descriptores relevantes. La implementación de dicha herramienta se confrontará con varios problemas históricos. Los problemas de procesar el lenguaje natural y de extraer información han sido atacados desde hace varias décadas [13], [2], [17]. Sin embargo, todas las investigaciones realizadas no han sido suficientes para construir un sistema que interprete el lenguaje natural con un rendimiento parecido al ser humano.

El lenguaje natural ha escapado a todos los esfuerzos de tratamiento computacional exhaustivo, al parecer debido a que el conocimiento lingüístico está asociado, de formas sutiles y desconocidas, con el conocimiento contextual que tiene el hablante [6] [17]. Este conocimiento contextual resuelve en muchos casos la ambigüedad que no puede resolverse en el ámbito lingüístico superficial. El problema con el conocimiento contextual es que se trata de cualquier conocimiento. Eso significa que en primer lugar debe existir una forma de describir el mundo y en segundo lugar que esa forma de describir el mundo tiene que asociarse de alguna manera con el conocimiento lingüístico específico.

Integrar el conocimiento lingüístico y el conocimiento del mundo para tratar el lenguaje natural ha sido y será el problema general de cualquier sistema de procesamiento del lenguaje natural (NLP - Natural Language Processing).

En el desarrollo histórico del tratamiento de este problema destaca la iniciativa de usar el NLP desde los inicios de la computación. Con el advenimiento de los computadores, se ha tenido la idea de usarlos para obtener sistemas de extracción de información rápidos e inteligentes [14]. Trabajos mucho más recientes como los de Lewis [12], nos indican que nuevas demandas en la recuperación de información proveen oportunidades al procesamiento del lenguaje natural para trabajar con técnicas ya probadas de recuperación estadística. La descripción compacta del contenido relevante de un documento puede incrementar la eficiencia de la clasificación del material textual como relevante y no relevante, pues justamente la selección de características es crítica en las tareas de clasificación. Las investigaciones en el área de recuperación de texto tienen evidencias que sugieren que las búsquedas de los usuarios finales, deberían ser en lenguaje natural en lugar de ser orientados al lenguaje controlado [12].

Un ejemplo de aplicaciones que requieren un NLP son las bibliotecas. Muchas de ellas tienen ciertamente un problema de almacenaje y de extracción de información. Las tareas como la catalogación y la administración en general, han sido controladas con éxito por las computadoras [14]. Aunque la automatización ha resuelto parte del problema, las actuales bibliotecas virtuales tienen que clasificar grandes volúmenes de contenidos de información. La cantidad de información ha aumentado, y su acceso exacto y rápido se ha hecho más difícil. Como consecuencia de esto la información relevante generalmente es difícil de encontrar, y es necesario una búsqueda exhaustiva por parte del usuario final.

Debido a esta necesidad se han creado estándares de recuperación y catalogación de datos bibliográficos [9] [3], los cuales deben ser aplicados a toda la información. El experto que realiza el proceso de catalogación debe conocer estos estándares. Además, requiere un conocimiento especializado del tema y una estrategia sistemática para catalogar correctamente un documento. Este proceso es inmanejable por su dimensión, si se considera el volumen de información y el tiempo que necesita un experto para realizarlo completamente. La solución parece simple: sustituir al experto con un sistema automatizado. Pero esto involucra varios problemas computacionalmente complejos: (1) procesamiento del documento en lenguaje natural, (2) conocimiento especializado del tema sobre el cual trata el documento y (3) conocer y aplicar una estrategia sistemática para catalogar correctamente un documento.

Al parecer la caracterización automática en la cual una herramienta de *software* intenta duplicar el proceso humano de la "lectura" es un problema muy ambicioso. Específicamente, la "lectura" implica extraer la información, sintáctica y semántica del texto, separar la relevante y descartar la irrelevante. Si se logra leer el texto se pueden automatizar los procesos de catalogación de documentos convenientemente. De esta manera, se pone de manifiesto la necesidad de un NLP [14].

Como se mencionó anteriormente, el procesamiento del lenguaje natural parece ser un problema común de muchas aplicaciones que manejan información y conocimiento [10]. Precisamente el objetivo de este trabajo es desarrollar una herramienta que procese texto en español y extraiga sus descriptores o palabras claves. La idea es que se tenga una representación resumida del documento, para que estos descriptores puedan servir a otras aplicaciones o módulos que procesen grandes volúmenes de datos y textos en lenguaje natural.

Estas aplicaciones pueden ser, por ejemplo, buscadores eficientes, herramientas de extracción de conocimiento, minería

de texto, catalogadores automáticos, etc., que utilizarán la salida de nuestra herramienta como fuente de información. Por ejemplo la asignación de descriptores temáticos a documentos de una biblioteca virtual es una característica que puede brindar utilidad a los usuarios finales, pues la búsqueda y catalogación temática se realizaría en base a dichos descriptores.

Esta fue una vista preliminar del problema y la solución que se persigue describir en este artículo. Esta propuesta no se aleja de los grandes esfuerzos de investigación que actualmente están dirigidos a desarrollar y mejorar las tecnologías de recuperación de información, recuperación de texto y documentos, búsquedas inteligentes y extracción de conocimiento. Todas estas tecnologías comparten el mismo problema: "el procesamiento del lenguaje natural por parte del computador". Este se ha convertido en uno de los objetivos que destaca en el futuro de la computación.

2 LA ESTRATEGIA A PROBAR

El resultado de investigaciones como las de *Wiebe, Hirst y Horton* [15], plantea que cualquier texto establece un contexto lingüístico sobre el cual las siguientes palabras deben ser entendidas. El escritor de un texto se dirige a un lector con el propósito de informar, divertir, colaborar en una tarea, etc. Los lectores deben inferir la intención subyacente como parte de su comprensión. En el artículo de *Wiebe* [15], se presentan recientes investigaciones sobre el uso del lenguaje en un contexto, y concluye que el uso del lenguaje involucra mucho más que creación y comprensión de palabras aisladas, por tanto las computadoras, al igual que las personas, deben alojar el contexto interpersonal y lingüístico si están usando el lenguaje de una manera natural.

Precisamente un grupo de estas investigaciones sobre el uso del lenguaje en un contexto, tratan de expresar el matiz y el estilo en el lenguaje [15]. La exacta escogencia de palabras, frases y estructura de las oraciones afectan el significado y el efecto preciso de una palabra. Un escritor elige (conscientemente o no) objetivos como por ejemplo ser formal o amigable, persuasivo o despectivo, claro u oscuro. Estos aspectos de una palabra son mucho más parte de su mensaje que su significado literal, y cualquier sistema de lenguaje natural sofisticado debe ser sensible a ello. Un problema particular son las "expresiones referidas", es decir las palabras o frases que un escritor usa para denotar algún objeto o entidad particular. Muchos de los trabajos en la generación del lenguaje buscan determinar las expresiones referidas en el contexto en que son expresadas.

Según los trabajos de *DiMarco y Hirst* [8], un escritor usa varias construcciones sintácticas con un objetivo estilístico 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 otra parte, el profesor *Joseph Williams*, de la Universidad de Chicago, sugiere algunas reglas de estilo que podrían ayudar a mejorar la claridad de la escritura [16]. Sus libros están dirigidos a los angloparlantes pero las reglas son lógicas y validas para el idioma español. Las recomendaciones de *Williams* son particularmente útiles debido a que están basadas en el lector. *Williams* reconoce que escribir claro es escribir de tal manera que el lector tenga claridad acerca de lo que esta leyendo. En la propuesta de *Williams*, a cada oración en un texto se le asocia un tópico y la secuencia de estos tópicos en un párrafo sirve para analizar su coherencia.

La idea de esta propuesta es utilizar las sugerencias de estilo que propone *Williams* [16]. Esto se puede realizar formulando una gramática simplificada y reglas de estilo con el fin de identificar los tópicos de un texto. Nuestra intención es usar los tópicos como información básica para extraer descriptores significativos de una colección de párrafos en un texto. Un modelo como el mostrado en la **Figura 1** ilustra el esquema planteado. Dicho esquema es un sistema de procesamiento de lenguaje natural que contiene un módulo de nivel gramatical (nivel morfológico y sintáctico) y un módulo de nivel interpretativo (nivel semántico y pragmático). Otro componente importante es la base de conocimiento, que permite relacionar el conocimiento lingüístico con el contexto (conocimiento del mundo).

Nivel Gramatical

Durante años se ha tratado de expresar las lenguas en forma de gramática, con el fin de reconocer los constituyentes lingüísticos y la estructura de las oraciones. Esto se ha realizado con diferentes teorías gramaticales. Además existen numerosos intentos por expresar la lingüística de una lengua con técnicas y modelos no simbólicos, tales como los

modelos probabilísticos y los conexionistas [1] [4] [13]. Los modelos conexionistas tienen limitaciones y solo han sido desarrollados en aplicaciones de reconocimiento del habla. Los modelos estadísticos por su parte, realizan el estudio de una muestra (corpus) para modelar el uso del lenguaje. Esto se hace con el fin de reducir la ambigüedad (al contar con las opciones más probables), y mejorar la eficiencia del sistema de NLP. A partir de un modelo estocástico se obtienen reglas gramaticales con probabilidades asociadas. Entonces independientemente del modelo, simbólico o estadístico, el procesamiento del lenguaje natural involucra alguna gramática.

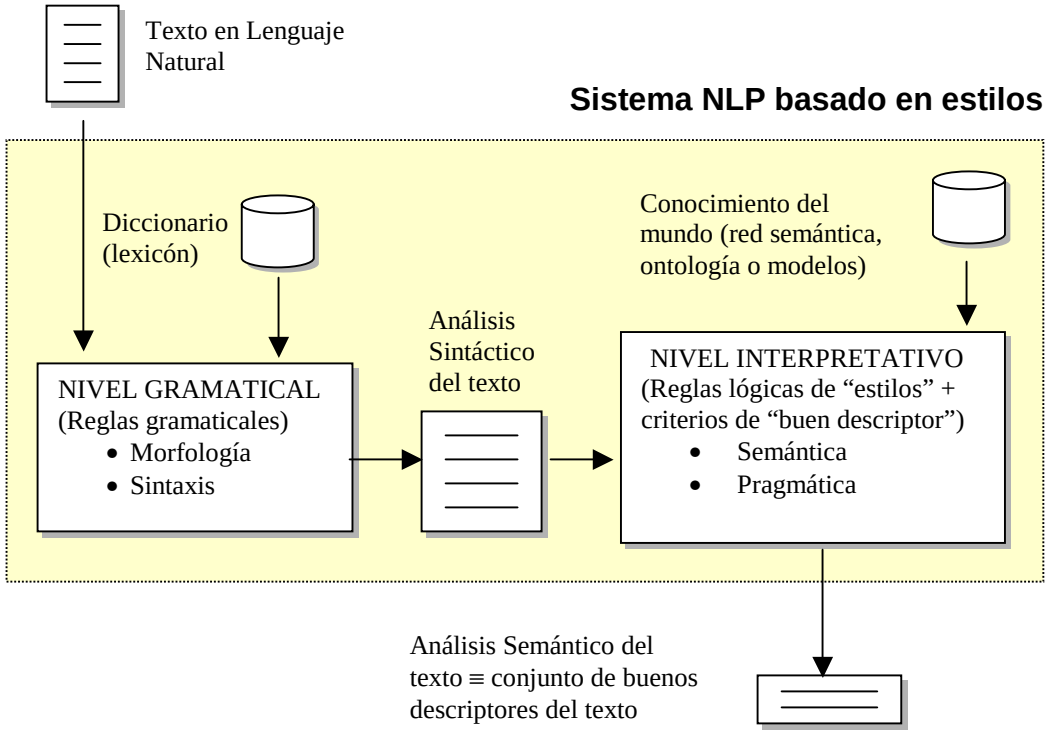

Figura 1. *Diagrama del procesamiento del lenguaje natural con una "gramática de estilos".*

Las gramáticas simbólicas no pueden incluir toda la información necesaria para tratar el lenguaje natural. Las gramáticas probabilísticas tienen el atractivo de incorporar tanto el conocimiento como el uso de la lengua; mientras que las gramáticas estándares solo manejan la competencia (conocimiento lingüístico). Estas limitaciones han creado la tendencia a desarrollar sistemas híbridos que combinan ambos esquemas.

Un sistema de NLP debe entonces resolver la interrogante: ¿cuál es la gramática más apropiada para describir formalmente las lenguas naturales? *Moreno* [13] afirma que muchas opiniones están de acuerdo en que la respuesta debe conjugar dos propiedades: (1) Expresividad: la gramática tiene que ser lo suficientemente poderosa como para que abarque todas las construcciones posibles en las lenguas naturales. (2) No sobre-generación: la gramática tiene que ser suficientemente restringida para que no permita como válidas construcciones agramaticales. Estas propiedades funcionan en la teoría, pero en la práctica las gramáticas formales se van modificando según las necesidades particulares: se introducen restricciones para reducir su poder o se incluyen extensiones para aumentar su expresividad. Las gramáticas expresan la sintaxis y la morfología y estas son diferentes entre los distintos idiomas. Por tanto, una gramática pierde la generalidad pues está atada a un léxico (vocabulario del dominio) y a un idioma.

Según *Moreno* [13] cualquier gramática computacional en español, de cierta cobertura, basada en la competencia, produce un número muy grande de análisis sintácticos alternativos para la mayoría de las oraciones. Sin embargo, es significativo que los hablantes no noten tanta ambigüedad cuando procesan oraciones: de manera espontánea solo tienen dos o tres posibilidades. Esto se debe a que la interpretación ayuda a eliminar las ambigüedades sintácticas, ya que los hablantes prefieren las interpretaciones plausibles a las poco probables. La plausibilidad de una interpretación viene

dada por el contexto semántico-pragmático. La solución para reducir el número de análisis y la ambigüedad sintáctica es incorporar restricciones semánticas y pragmáticas muy finas. Esto es una manera de introducir el conocimiento del mundo a un sistema de NLP.

Para este trabajo de investigación se debe definir formalmente una gramática que use un procesamiento lingüístico simplificado, básicamente sintaxis. El fin de esta gramática no es realizar una identificación exhaustiva de los constituyentes gramaticales de la oración. Su auténtico objetivo es identificar solo aquellos constituyentes que nos permitirán luego aplicar las reglas lógicas del nivel interpretativo de nuestro sistema. Para esto se necesita identificar especialmente el verbo de la oración, a través de un sistema de conjugación verbal para el español. Los verbos deben estar clasificados, los irregulares deben colocarse todos. Podemos identificar al resto de los verbos clasificados (según su comportamiento similar) usando morfología.

El *parser* o analizador de la estructura gramatical del lenguaje, debe implementar la gramática definida formalmente en el paso anterior. Se debe usar un lenguaje declarativo, como Prolog, para evitar desviar la atención hacia el procesamiento de las reglas gramaticales.

Nivel Interpretativo

El conocimiento del mundo es la representación del mundo real del hablante y de su dominio de conocimiento. Este conocimiento del mundo interviene en el nivel interpretativo del lenguaje, también llamado la semántica. Este nivel interpretativo es más general que el nivel lingüístico, pues parece no depender directamente del idioma. Por ejemplo *Williams* [16] plantea reglas de estilos y estas pueden ser aplicadas no solo al idioma inglés, para el cual fueron concebidas, sino también a otros idiomas. Así se pueden tener varios niveles interpretativos del lenguaje: el primero es la semántica de las oraciones aisladas. Esta semántica oracional puede ser expresada de manera lógica, por ejemplo "Don Quijote ama a Dulcinea" se representa lógicamente como la cláusula `ama(Don Quijote, Dulcinea)`. El segundo nivel es la semántica del discurso, relación de las oraciones entre si, la cual se puede simbolizar también a través de reglas lógicas.

El conocimiento del mundo se ha logrado expresar con técnicas de representación del conocimiento, tales como modelos, redes semánticas y ontologías. El conocimiento del mundo del hablante se puede entender entonces como una base de conocimiento. Este conocimiento del mundo interviene en la interpretación oracional y del discurso, pues sobre la base de dicho conocimiento se maneja el conocimiento implícito de los hablantes, que permiten resolver las ambigüedades, y también se resuelven las *anáforas*[1] y *elipsis*[2]. El nivel interpretativo se puede definir como un conjunto de reglas lógicas que procesan la información gramatical (sintáctica) y el conocimiento del mundo.

Entonces se puede pensar que un procesamiento lógico permitiría lograr un nivel interpretativo en una gramática. Las reglas lógicas para la interpretación oracional y del discurso, pueden complementar las reglas gramaticales. El mismo nivel interpretativo, basado en reglas lógicas, podría procesar la representación del conocimiento del mundo que debe tenerse para el dominio de conocimiento.

Como se mencionó antes, lo que hasta ahora se piensa es realizar una gramática que permita identificar los constituyentes básicos de la oración. Complementar esta gramática con las reglas de estilo de *Williams*, para darle la semántica adecuada al texto y así identificar las acciones y los actores. Esto puede permitir identificar los descriptores, palabras claves o tópicos con los cuales obtener una descripción breve del texto.

Como dice *Rijsbergen* [14], es intelectualmente posible que un ser humano determine el contenido relevante de un documento. Para que un computador haga esto se necesita realizar un modelo dentro del cual se definan las decisiones de relevancia. Es interesante observar que la mayoría de las investigaciones sobre la recuperación de información muestran haber tratado diversos aspectos de tal modelo [14]. De esta manera, otro componente importante de este nivel interpretativo son los criterios para definir un "descriptor relevante" o "buen descriptor".

[1] La anáfora se refiere a una expresión previa de un discurso en lenguaje natural. Generalmente usa un pronombre para referirse a personas, lugares o cosas previamente mencionadas. Por ejemplo "María murió. Ella estaba muy vieja", ella se refiere a María.

[2] La elipsis se refiere a las situaciones cuyas oraciones son abreviadas o eliminan un constituyente, dejando parte de ellas para ser entendidas por el contexto. Por ejemplo cuando se pregunta "¿Cuál es tu nombre?", y se contesta "Juan Pérez" esta es una forma elíptica de "Mi nombre es Juan Pérez".

Un descriptor relevante para nosotros sería una consecuencia lógica del discurso que estemos analizando:

$$\texttt{Discurso} \models \texttt{descriptor}$$

En este caso no conviene cualquier consecuencia lógica (pueden haber muchas). Se escogerá una de las que mejor resuma el texto:

$$\texttt{representación-del-discurso} \cup \texttt{teoría-de-apoyo} \models \texttt{buen\_descriptor(descriptor)}$$

La teoría de apoyo es una axiomatización que nos dice que ciertas frases de ciertos discursos son o no buenos descriptores de esos discursos. La representación del discurso contiene al discurso y algunas ideas acerca de su contenido. Aquí es donde intervienen las reglas de estilo de *Williams*, debido a que usaremos sus reglas para asociar tópicos al discurso.

Esto va a permitir tener criterios para determinar el éxito y el fracaso en la extracción de información relevante. Es decir, cómo se puede distinguir del conjunto de los descriptores, aquellos que son candidatos a ser catalogados como buenos. Algunas alternativas que pueden tomarse en cuenta en esta tarea pueden ser: técnicas de modelos probabilísticos para detectar relevancia en el contexto [4], redes semánticas, ontologías del dominio de conocimiento y criterios bibliográficos para asignar descriptores a documentos, por ejemplo CEPAL (Comisión Económica para América Latina y el Caribe)[3].

3 REGLAS DE ESTILO

El profesor *Joseph Williams*, de la Universidad de Chicago, sugiere algunas reglas de estilo que podrían ayudar a mejorar la claridad de la escritura. Las recomendaciones de *Williams* son particularmente útiles debido a que están basadas en el lector. Sus recomendaciones comienzan en el ámbito de la oración. *Williams* cree que los lectores encuentran a las oraciones fáciles de leer y entender cuando la lógica del pensamiento 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. 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.

Su ayuda continúa a nivel del párrafo. Las oraciones que constituyen un párrafo deberían tener tópicos consistentes. 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.

Williams presenta elementos que permitirán dar fundamento a ciertas reglas en los documentos escritos. Estos elementos indispensables para obtener un estilo de escritura legible, son claridad, cohesión, énfasis, coherencia, elegancia, concisión y longitud [16]. Los primeros tres son los más útiles y mejor planteados por *Williams* [7].

Claridad

La claridad nos permite identificar de manera precisa los actores y acciones del relato. *Williams* enuncia los primeros principios de la escritura clara:

1. Los sujetos de las oraciones enuncian el reparto de personajes.
2. Los verbos que van con estos sujetos enuncian las acciones cruciales de los cuales aquellos personajes son parte.

Para cumplir con estos principios es suficiente usar una forma clara y transparente al presentar los verbos y los sujetos. Una manera común de ocultar los actores es usar la “voz pasiva” y la “sustantivación”[4]. Sin embargo, el uso de la sustantivación puede ser adecuado y válido en ciertos momentos, *Williams* se refiere a la tendencia de abusar de su uso,

[3] La definición de descriptor temático según CEPAL es la siguiente: "Términos formados por una o más palabras claves que resumen o denotan un concepto, extraídos de un tesauro o vocabulario controlado utilizado por la unidad de información".

[4] La sustantivación es un recurso del lenguaje que consiste de la transformación de acciones en sujetos. Se usa para ocultar los verbos (acciones). También existe cuando se transforma un adjetivo en nombre (nominaliza un adjetivo). Por ejemplo el verbo “descubrir”, es substantivado con la palabra “descubrimiento”.

lo cual va en detrimento de la claridad de la escritura. Para evitar esto y ser claros, es necesario saber que los lectores identifican dos niveles de estructura dentro de una oración o frase la cual debemos integrar:

(1) su secuencia gramatical predecible: sujeto + verbo + complemento
(2) su historia, un nivel de significado cuyas partes tiene un orden fijo: caracteres + acciones.

Estructura fija	Sujeto	Verbo	Complemento
Contenido variable	Caracteres	Acciones	----

Cuadro 1: *Claridad*

Los elementos fijos aparecen en toda oración o frase que posea un sentido completo. Se trata de una estructura fija a la cual debemos tratar de asignar agentes y acciones. Los elementos variables, por su misma condición, se pueden mover de cualquier manera o incluso pueden no aparecer dentro de la oración. Lo que propone *Williams* es hacer coincidir los caracteres con el sujeto y las acciones con los verbos.

El personaje es un elemento variable del nivel histórico de la oración. En general, hay muchos tipos de personajes. Los más importantes son los agentes -el origen directo de cualquier acción o condición-. Es decir, son los causantes de las acciones. Es importante destacar que cuando se hace coincidir el personaje con el sujeto de la oración, éste se convierte en el agente, es decir, el causante directo de la acción.

Las acciones también se identifican como elementos variables de la estructura histórica de la oración. Entendemos por acciones a aquellas palabras que describen un estado de alteración (física, psicológica o espiritual) del personaje respecto a su ambiente. En contraste entendemos por verbos a aquellas acciones que afectan, no al personaje, sino al agente de la oración.

Cuando se trata de interpretar un párrafo completo es importante conocer el concepto de “cadena lógica consiste de sujetos” [16]. Esto permite que el lector identifique los agentes de las acciones de cada una de las oraciones, realizando conexiones lógicas entre ellas. La idea es que al inicio se logre “anclar” al lector en un concepto que le es familiar o conocido para introducir luego un concepto nuevo. Otra recomendación es usar al principio una oración que oriente al lector en que es lo que sigue. La cadena consistente de sujetos debe corresponder con los argumentos de una inferencia lógica.

Lo anterior se refiere a la claridad local interna de cada oración independientemente de un contexto o de una intención. Existen dos elementos mas, a parte de la claridad, que deben considerarse para alcanzar un texto legible. Ellos son la cohesión y la coherencia.

Cohesión y Coherencia

Se entiende por Cohesión a “la manera como las diversas oraciones que conforman un texto escrito permanecen unidas bajo un mismo contexto o discurso” [16]. Con cada oración que escribimos debemos establecer el mejor encuentro entre los principios de la claridad local y los principios de cohesión que unen oraciones separadas dentro de un mismo discurso.

Williams ofrece dos principios para mantener la cohesión:

(1) Colocar al inicio de una oración aquellas ideas que ya hayan sido enunciadas, referidas e implicadas, o bien aquellos conceptos que pueden asumirse como familiares y conocidos por el lector.
(2) Colocar al final de la oración lo más nuevo o reciente, lo más sorprendente, la información más significativa, es decir la información que desea extender y desplegar.

El “tópico” es un concepto que juega un papel preponderante en la búsqueda de la cohesión. *Williams* define el tópico como "el sujeto psicológico de la oración". Es decir, el tópico es el elemento que lleva la carga lógica del discurso, tanto oral como escrito. Desde el punto de vista lógico los tópicos son los conceptos emitidos o involucrados en cada una de las proposiciones que posee un argumento, ya sea en sus premisas o en su conclusión.

El tópico es casi siempre una frase nominal de cualquier tipo que el resto de la oración caracteriza o comenta, sobre la que se dice cualquier cosa. Son ideas que definen de que se trata el texto. En forma incremental estas ideas topicalizadas

proporcionan avisos temáticos que enfocan la atención del lector hacia un conjunto bien definido y limitado de ideas conectadas.

Las cadenas lógicas y consistentes de sujetos (claridad) tienden a solaparse con las cadenas de tópicos (cohesión). Ambas cuentan con un criterio similar en su construcción, y serán la misma en la medida que nuestros tópicos sean los sujetos de nuestras oraciones. Sin embargo, cuando existan puntos donde estos dos criterios sean divergentes, debe prevalecer el criterio de cohesión.

Gráficamente se puede representar la cohesión como sigue, manteniendo la consistencia con la representación de la claridad:

Estructura fija	Tópico	Énfasis
Contenido variable	Vieja inf. familiar	Nueva inf. no familiar

Cuadro 2: *Cohesión*

En el **Cuadro 2** se observa que la vieja información y la nueva información poseen relación entre sí. A esta relación es lo que *Williams* llama ideas topicalizadas. Es decir, la nueva información esta referida al tópico que ha sido mencionado en la vieja información.

Entendemos por vieja información, aquellos conceptos contenidos en las oraciones que son conocidos por el lector y donde potencialmente encontraremos al tópico. En este caso, la vieja información es análoga a personajes, y el concepto seleccionado como tópico es análogo a agente.

Entendemos por nueva información a aquellas oraciones que poseen nuevos y más complejos conceptos. Estos poseen nuevos tópicos que nos permitirán movernos, "lógicamente", entre diversos contextos asegurando así la cohesión del discurso escrito.

Por otra parte, *Williams* define párrafo cohesivo como aquel que tiene una cadena consistente de tópicos (denominadas cadenas temáticas). Un párrafo cohesivo induce un nuevo tópico y cadena temática en una posición predecible, al final de la(s) oración(es) introductoria(s) del párrafo. Este principio nos permite identificar dos nuevos elementos dentro del párrafo: "salida o arranque" y "discusión". Es decir, que un párrafo coherente tendrá usualmente una oración sencilla que claramente articule su tema o idea principal. De la misma manera un párrafo coherente ubicará típicamente aquella idea principal o punto[5] en uno de los dos lugares siguientes: en el arranque o en la discusión del párrafo.

Williams también establece qué hacer cuando hay conflicto entre las reglas. En el modelo de *Williams* existen prioridades donde la coherencia prevalece a la cohesión y esta a su vez al principio de la claridad.

Según *Williams,* los elementos básicos de una escritura legible que encuentra un lenguaje útil en la comunicación son la claridad, cohesión y coherencia. Porque logran una calidad en la información, así como una comunicación clara, precisa y efectiva.

4 UN EJEMPLO DE LA ESTRATEGIA

Para ilustrar la estrategia, se utiliza como discurso el texto de una noticia internacional, tomada de *"The Wall Street Journal Americas"* (15-02-2001). A este texto se le aplica una serie de reglas propuestas para extraer los descriptores. Estas reglas están basadas en las reglas de estilo de *Williams* y en un posible criterio de descriptor relevante. En este caso, el criterio de descriptor relevante es el "tópico común más específico"[6] en todo el texto.
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:

[5] *Williams* llama "punto" a la idea principal de un documento.
[6] Denominamos "tópico común más específico" a aquel tópico que es menos general. La generalidad la determinamos en función de la longitud de la cadena de palabras que constituyen el tópico, por lo tanto un tópico más general es más breve.

(1) T es un tópico del discurso D si en el discurso D, un Agente *revela* T
(2) 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.
(3) 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* Agente".
(4) 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.
(5) T' es más general que T si T'es más breve que T.

Las reglas (1), (2) y (3) expresan las recomendaciones de estilo de *Williams* donde se identifican los agentes, las acciones (verbos) y se extraen los tópicos. Las reglas (4) y (5) contienen el criterio para determinar la relevancia del descriptor. Aplicándolas todas se obtiene como resultado el tópico T = **"mal de las vacas locas en cabras y ovejas**".

5 EVALUACIÓN DEL SISTEMA

Para la evaluación de esta estrategia es pertinente diseñar varios experimentos para obtener resultados que puedan ser validados. Por supuesto, estos experimentos deben coincidir con el dominio de conocimiento escogido. Es importante permitir la interacción de un experto en el dominio del contenido del texto para comparar la relevancia obtenida por el sistema. Se tendrá por lo menos dos grupos de control: (1) Grupos de descriptores determinados por expertos y (2) Grupos de descriptores determinados por expertos pero asistidos por la herramienta.

Debemos aplicar un método convencional de evaluación en donde tendríamos que examinar el concepto de relevancia, la cual es una noción subjetiva. Los humanos pueden diferenciar sobre la relevancia o la no-relevancia del contenido de los documentos. Esta noción de la relevancia ha sido explicada por *Cooper* [5] y la denomina correctamente "relevancia lógica". La importancia esta 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.

Sin embargo, lo que se quiere es tener una estrategia general para extraer los descriptores relevantes independientemente de una necesidad específica de información. Es decir, los descriptores pueden ser relevantes o no dependiendo de la utilidad que se le dará, la cual es diferente para efectos de catalogación, terminología, procesamiento lingüístico o búsqueda de información. Una vez definido la utilidad del descriptor, será mas sencillo definir la relevancia.

En vista de la importancia de la evaluación, será necesario considerar en este trabajo algunas metodologías de evaluación de sistemas NLP, aunque estos métodos pueden resultar muy costosos y complejos. *Margaret King* [11] destaca la importancia de los resultados de las evaluaciones a sistemas de NLP como datos invalorables, pero advierte que las evaluaciones varían enormemente en función del propósito, del alcance y de la naturaleza de los objetos que están siendo evaluados.

CONCLUSIONES

El problema planteado en esta propuesta consiste en definir una estrategia para interpretar textos escritos en español y extraer buenos descriptores, además de explorar criterios lógicos para asociar descriptores adecuados y relevantes a textos en español.

Este problema requiere un procesamiento del lenguaje natural que puede necesitar el conocimiento lingüístico del idioma y el conocimiento del mundo que maneja el hablante. El conocimiento lingüístico, en particular la morfología y sintaxis, puede ser expresado mediante formalismos lingüísticos que permitan reconocer los constituyentes y la estructura de las oraciones. El conocimiento del mundo puede ser implementado con (1) reglas lógicas inspiradas en los estilos de *Williams*, (2) criterios de descriptor relevante y (3) representación y uso de un dominio de conocimiento.

De esta manera, se tiene como hipótesis que, si se aplica una gramática de estilos, entonces pueden obtenerse descriptores relevantes de un documento escrito basado en estos estilos. Se espera que los descriptores derivados de esta estrategia sean próximos a los descriptores obtenidos por los expertos en el dominio.

Este proyecto puede realizar varios aportes. El primer aporte, básicamente teórico se refiere al modelo de la gramática de estilos basada en las recomendaciones de *Williams*. Incorporar estas reglas de estilo para la semántica al análisis gramatical tradicional pudiese ser muy efectivo para asociar resúmenes a textos de un dominio específico.

En segundo lugar podría desarrollar una herramienta que interprete documentos en español y extraiga de ellos descriptores relevantes, es un aporte práctico muy útil, debido a que representan una descripción compacta del contenido de un documento. Esta descripción puede incrementar la eficiencia en diferentes tareas y herramientas como minería de texto, catalogadores automáticos, buscadores eficientes, etc., de manera que el grupo de descriptores puede reemplazar la búsqueda exhaustiva sobre todo el texto.

Como en cualquier ciencia aplicada, hay cierta separación entre lo esperado teóricamente y los resultados prácticos. El funcionamiento real de los sistemas NLP depende de muchas variables que generalmente contradicen las demostraciones teóricas. Entonces, para determinar el éxito o no de esta propuesta, se debe considerar cuales son estas variables y su influencia, además del uso práctico e importancia teórica de la aplicación desarrollada.

REFERENCIAS

[1] Abney, S. "Statistical Methods and Linguistics", en Klavans y Resnik (1996).
[2] Allen, J. "Natural Language Understanding". Redwood City. Benjamin/Cummings. (1995).
[3] Chan, L.M., Comaromi, J.P., Mitchell, J.S., Satija, M.P. "Dewey Decimal Classification: A Practical Guide", 2nd edition, OCLC Forest Press, Albany NY, (1996).
[4] Charniak, E. "Statistical Language Learning", Cambridge, The M.I.T. Press. (1993).
[5] Cooper, W.S. "A definition of relevance for information retrieval", Information Storage and Retrieval, 7, 19-37 (1971).
[6] Covington, Michael A. "Natural Language Processing for Prolog Programmers". Artificial Intelligence Programs The University of Georgia Athens, Georgia. PRENTICE HALL, Englewood Cliffs. NewJersey 07632 (1994).
[7] DeLong, J. Bradford. "Review of Joseph Williams, Style: Toward Clarity and Grace". December (1999). http://www.j-bradford-delong.net/Econ_Articles/Reviews/Williams.html
[8] DiMarco, C. y Hirst, G. "A computational theory of goal-directed style in syntax", Computational Linguistics. 19, 3, 451-499, (Septiembre 1993).
[9] ISO 23950. "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)" Bethesda, MD: NISO Press. 180pp. ISBN: 1-880124-22-X ISSN: 1041-5653. (1995).
[10] Jacobs, Paul y Rau, Lisa. "Innovations in text interpretation". Artificial Intelligence 63. (1993).
[11] King, M. "Evaluating Natural Language Processing Systems", Communications of the ACM., Vol. 39, No. 1. (Enero 1996).
[12] Lewis, David D. y Sparck Jones, Karen. "Natural Language Processing for Information Retrieval", Communications of the ACM., Vol. 39, No. 1. (Enero 1996).
[13] Moreno Sandoval, Antonio. "Lingüística Computacional. Introducción a los modelos simbólicos, estadísticos y biológicos". Madrid. Editorial Síntesis. (1998).
[14] Rijsbergen, C. J.. "Information Retrieval". Second Edition (London: Butterworths, 1979).
[15] Wiebe, J., Hirst, G. y Horton, D. "Language Use in Context", Communications of the ACM. Vol. 39, No. 1. (1996).
[16] Williams, Joseph. M., "Style: Toward Clarity and Grace". The University of Chicago Press. Chicago and London. (1990).
[17] Winograd, T. "Language as a Cognitive Process: Syntax". Reading, Addison-Wesley. (1983).
[18] Winograd, T. "Understanding Natural Language", Edinburgh University Press, Edinburgh (1972).