Metodología para determinar la exactitud de una respuesta, escrita en forma textual, a un interrogante sobre un tema específico, aplicando herramientas informáticas.

María Alejandra Paz Menvielle, Mario Alberto Groppo, Marcelo Martín Marciszack, Analía Guzmán, Karina Ligorria, Martín Cassatti, Seiyu Ricardo Higa Tamashiro, Juan Pablo Gimenez

Departamento de Ingeniería en Sistemas de Información
Facultad Regional Córdoba – Universidad Tecnológica Nacional

pazmalejandra@gmail.com, proyale@groppo.com.ar, marciszack@gmail.com, aguzman@sistemas.frc.utn.edu.ar, karinaligorria@hotmail.com, mcasatti@gmail.com, rickyssht@gmail.com, gimenezjuan92@gmail.com

Resumen

El presente proyecto realizará análisis de texto en un párrafo, de redacción libre, que es respuesta a preguntas relacionadas a un dominio específico, con el objetivo de detectar si esa respuesta es correcta. Las pruebas se realizarán en una cátedra, donde sus alumnos responderán a preguntas de exámenes usando texto de redacción libre. La cátedra es Paradigmas de Programación, 2do año de Ing. en Sistemas de Información en UTN FRC, ya que la misma posee contenidos con dominios simples y directos que facilitan la interpretación de las posibles respuestas de los alumnos. Para el análisis se definirá el dominio de aplicación tanto en relación a la temática que se trate como en la forma en que el alumno deba dar la respuesta. Se estudiará la necesidad de establecer para las respuestas ciertas restricciones que facilitarían su posterior estudio.

Palabras clave: análisis de textos, grafos, análisis sintáctico de textos, análisis semántico de textos

Contexto

El presente trabajo forma parte del proyecto de investigación y desarrollo que ha sido homologado por la Secretaría de Investigación, Desarrollo y Posgrado de la Universidad Tecnológica Nacional, reconocido con el código PIDEIUTNCO0003592 en el ámbito de la Universidad, por un período de dos años y a partir del 1 de enero de 2015.

Debe dar cumplimiento en simultáneo a los requisitos fijados para la asignatura y cumplir con los descriptores y criterios de intensidad de formación práctica de la Resolución Ministerial 786/09.

Introducción

El tratamiento de textos con diferentes finalidades, es tema de estudio en diversas disciplinas como medicina, ingeniería, biología, etc. En todos los casos se emplean metodologías originadas en otras áreas tales como extracción de información, recupero de información, lingüística computacional, minería textual, categorización de textos, etc. También se han realizado numerosos trabajos relacionados con el análisis de textos, como son la generación automática de resúmenes y los sistemas de búsquedas de respuestas, entre otros.

El equipo de trabajo cuenta con docentes pertenecientes a la cátedra Sintaxis y Semántica de Lenguajes (en la que se

estudian las gramáticas formales y lenguajes estructurados), y docentes y alumnos especializados en programación y en el estudio de los grafos y sus aplicaciones, que harán posible el desarrollo del este proyecto.

El estudio de los lenguajes estructurados

El orden lineal de las frases, como su estructura jerárquica, son el tema principal en los formalismos para el análisis sintáctico. Los distintos enfoques consideran esa jerarquía como relaciones entre combinaciones de las palabras o entre palabras mismas.

En el Paradigma de Chomsky se han desarrollado muchos formalismos para la descripción y el análisis sintácticos. El concepto básico de la gramática generativa es simplemente un sistema de reglas que define de una manera formal y precisa un conjunto de secuencias (cadenas a partir de un vocabulario de palabras) que representan las oraciones bien formadas de un lenguaje específico.

Chomsky (1965) y posteriormente sus seguidores desarrollaron y formalizaron una teoría gramatical basada en la noción de generación. El trabajo que se realiza en la gramática generativa descansa en la suposición acerca de que la estructura de la oración está organizada jerárquicamente en frases (y por consiguiente en estructura de frase) [1][2], y éste es el punto de partida de nuestro estudio.

Generación automática de resúmenes

En [3] se define el problema de la generación de resúmenes automáticos a dos diferentes niveles: a nivel de documento y a nivel de grupos de documentos. Un problema común es el de la existencia de múltiples documentos sobre un mismo tema, en este caso se habla de resúmenes a nivel de colecciones de documentos que agrupan o separan los documentos por tópicos y destacan las similitudes y diferencias de la información contenida en ellos. Los contenidos se relacionan entre ellos en un sentido semántico: cubren el mismo tópico y tienen similares categorías semánticas o conceptos estrechamente relacionados.

Categorización de textos

Dado un grupo de objetos, la tarea de clasificarlos consiste en asignarlos a un conjunto pre especificado de categorías. Si estamos dentro del dominio de gestión documental, la tarea se la conoce como categorización de texto, y consiste en hallar uno o más tópicos en los que encajen los contenidos de los documentos; teniendo como entrada un grupo de categorías (sujetos – temas) y un conjunto de documentos de texto. La categorización automática de documentos es una forma de clasificación de patrones, que es necesaria para la gestión eficiente de sistemas de información de texto. Se aplica en el indexado de texto para entrega comercial personalizada de texto, filtrado de spam, categorización de páginas web bajo catálogos jerárquicos, generación automática de metadatos, detección de género de textos, etc. [4].

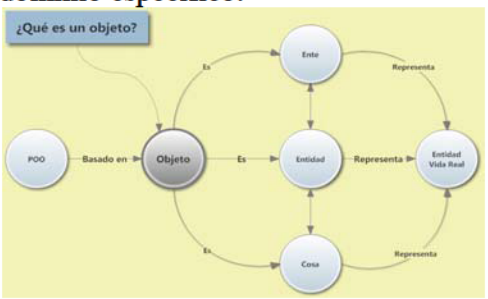
Sistemas de búsquedas de respuestas

En [6] se describen los sistemas de búsqueda de respuestas como sistemas diseñados para tomar una pregunta en lenguaje natural y entregar una respuesta. De esta manera los usuarios no tendrían que navegar y leer una o varias páginas de resultados de búsqueda. Estos sistemas se construyen sobre motores de búsqueda y requieren contenido como fuente para descubrir las respuestas. Deben tener métodos para entender las preguntas del usuario y determinar el tipo de respuesta que debe dar, generar una búsqueda significativa de la consulta, y finalmente calificar los resultados obtenidos. De estos tres problemas el más difícil de enfrentar es determinar el tipo de

respuesta. Para hacerlo se realizan tres pasos: entrenamiento, fragmentación y solo entonces la determinación del tipo de respuesta. Para construir un sistema de respuestas se necesita aplicar técnicas de PLN como: Recupero de Información, Extracción de Información con algoritmos de reconocimiento de entidades y etiquetado, comparación de secuencias de caracteres, entre otras.

Grafos

La teoría de grafos (Graph Theory), Dieter (2004), es ampliamente utilizada. En el presente proyecto se evaluará la aplicación de grafos para establecer la relación entre los conceptos que conforman el conocimiento dentro de un dominio específico.



Líneas de Investigación, Desarrollo e Innovación

Para el desarrollo de este proyecto de investigación se utilizará el método científico, con la aplicación de un desarrollo experimental, el cual se implementará en dos años.

En el primer año se continuará con la investigación teórica y de las herramientas existentes sobre análisis de textos, se definirá el dominio de aplicación y se comenzará con el diseño del motor de corrección, el núcleo, que permite establecer la base de conocimiento, definir preguntas, analizar respuestas y obtener un grado de validez para dichas respuestas.

Durante el segundo año se finalizará el diseño del sistema y se desarrollará la aplicación. Paralelamente a esta tarea, se diseñará el plan de pruebas que incluirá la elaboración de exámenes modelo y sus posibles respuestas Se iniciará la implementación del sistema con algunos docentes seleccionados quienes deberían participar, esporádicamente y como consultores, para definir la forma de elaborar las preguntas. Se requerirá la participación de alumnos de la cátedra Paradigmas de Programación, cuyas respuestas a las preguntas servirán para evaluar y analizar los resultados obtenidos. Se realizará un seminario de capacitación destinado a los docentes de la cátedra para mostrar la funcionalidad de la herramienta construida.

Al finalizar el proyecto, se elaborarán informes comparativos y conclusiones sobre el uso integrado de las herramientas implementadas y se realizará la comunicación de los resultados a la comunidad universitaria.

Resultados y Objetivos

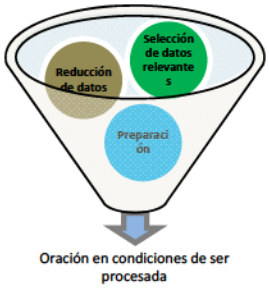
El presente proyecto es continuidad de trabajos anteriores realizados en el seno de la cátedra Paradigmas de Programación y de la investigación realizada en el marco del proyecto de tesis de uno de los integrantes del grupo de investigación; actualmente se está trabajando en el estudio, técnicas y aplicaciones de software existentes, que permiten realizar análisis sintáctico de oraciones, como también correctores ortográficos.

Tareas clave:

La tarea más importante y de mayor complejidad es el pre-procesamiento y transformación del texto respuesta. Es necesario transformar los datos que se encuentran en las respuestas a las preguntas, en datos aptos para realizar el proceso de análisis.

A este proceso se lo puede subdividir en:

a. Selección de datos relevantes: busca eliminar entradas duplicadas o redundantes y detectar anomalías en el conjunto de datos.
b. Reducción de datos: seleccionar las características adecuadas para el proceso de análisis de los datos
c. Preparación: realiza la limpieza, integración, transformación y reducción de dimensiones sobre los datos seleccionados.



Es importante esta etapa ya que los resultados obtenidos dependerán en gran parte de la calidad de los datos con los cuales se trabaja.
En relación al procesamiento del texto, durante la investigación preliminar surge la necesidad de realizar el tratamiento en diferentes niveles: desde el nivel de palabra a palabra, frase (no usada realmente en la actualidad), respuesta, conjunto de respuestas, conjunto de respuestas con enlaces y aplicación. Cada nivel de procesamiento revela diferente información sobre el texto.
Este procesamiento es la clave del proyecto. Es importante optimizar la resolución del problema y la solución deberá provenir desde la presente investigación. No es sencillo, un lenguaje consiste en un alfabeto, una gramática y un conjunto de reglas que definen la sintaxis. El alfabeto es el conjunto de símbolos usados por un lenguaje. La gramática de un lenguaje es un conjunto de reglas que definen cómo los símbolos de un alfabeto pueden interactuar uno con otro, mientras la sintaxis consiste en reglas que capturan el modo que las palabras pueden ser unidas para formar una sentencia, véase [5] y [10]. Todas las "fugas de las gramáticas" suceden porque la gente tiende a usar el lenguaje libremente, sin adherirse a las reglas. Describir texto partiendo sólo de la gramática puede conducir a identificaciones erróneas, a inhabilidad para capturar errores sintácticos en el texto o para identificar ciertos ítems tales como nombres, aunque las reglas sintácticas básicas puedan capturar patrones claves en la estructura del lenguaje.
Con el objeto de eliminar los errores de interpretación, comunes en las respuestas de texto libre, el enfoque de este trabajo requiere establecer ciertos acuerdos en cuanto a la gramática utilizada en la formulación de preguntas y respuestas.
El objetivo principal del presente proyecto es detectar patrones de texto en las respuestas, escritas en forma de texto libre, a preguntas dentro de un dominio específico, para determinar si lo escrito responde acertadamente, o el grado de proximidad a una respuesta correcta.
Para ello se construirá una herramienta de software que permita cargar las respuestas a las preguntas, analizarlas en base a la gramática redefinida, e indicar el grado de exactitud de las mismas.

Formación de Recursos Humanos

Los resultados del presente proyecto serán un importante aporte para los docentes de distintas cátedras (por ejemplo Sintaxis y Semántica de Lenguajes), ya que enriquecerá a la comunidad académica en su conjunto.
Del mismo modo el conocimiento obtenido y las conclusiones que se logren, en relación al procesamiento de textos, así como a la aplicación de los paradigmas de programación, aportarán al

enriquecimiento en la formación de los integrantes del equipo de investigación.
En el equipo trabajarán estudiantes de la Carrera de Ingeniería en Sistemas de Información, con la finalidad de que inicien su formación en investigación científica y tecnológica, profundizando sus conocimientos en temas significativos de la especialidad. Los estudiantes tendrán la posibilidad de hacer la Práctica Supervisada, de quinto año, en el marco del proyecto.
También se incorporará a docentes investigadores de la carrera que comenzarán a capacitarse en los procesos de investigación, y a un egresado de la Carrera de Ingeniería en Sistemas de Información, con la finalidad de que inicie su formación en investigación científica y tecnológica, y pueda aportar a la comunidad los conocimientos adquiridos.

Además está planteado el desarrollo de las respectivas Tesis de Maestría de dos integrantes docentes del presente proyecto.
El análisis de textos desde esta perspectiva continuará ampliando el aprendizaje sobre la materia y las distintas técnicas y metodologías existentes, permitiendo de esa manera acercar a la comunidad académica un poco más a la solución de la problemática permanente que significa que un computador pueda analizar textos correctamente.
El producto de este trabajo puede ser llevado, en el marco de transferencia tecnológica, a otros entornos que requieran un análisis automático de respuestas, por ejemplo Mesas de Ayuda de empresas u organismos públicos, consultas a bases textuales, etc.
La capacidad de poder analizar automáticamente una consulta o pregunta elaborada por una persona sin trasfondo técnico mejoraría dramáticamente la interacción entre los sistemas de almacenamiento de información digitalizada y el público en general.

Referencias

[1] "Teoría de la Computación", J Glenn Brookshear. Ed Addison Wesley – 1989.
[2] "Teoría de Autómatas, Lenguajes y Computación". J. Hopcroft, R, Motwani, J. Ullman. Ed. Pearson- Addison Wesley – 2008.
[3]- Aplicaciones de Procesamiento de Lenguaje Natural. Hernández M._ Gómez J. Revista Politécnica, Julio 2013, Vol. 32, No. 1, Páginas: 92–96
[4] "The text mining handbook", R. Feldman and J. Sanger. Cambridge University Press. Diciembre 11, 2006. También disponible on line: www.safaribooksonline.com
[5] "Lenguajes, Gramáticas y Autómatas, un enfoque práctico". Isasi, P., Martines, P., Borrajo, D. Ed Addison-Wesley. 1997.
[6] "Introduction to information retrieval". C. D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Cambridge University Press. Julio 2008. También disponible on line: www.safaribooksonline.com
[7] "Minería de texto para actualización automática de documentos"Advanced Approaches to Analyzing Unstructur. Pérez Abellería, M.A. y Cardoso, C.A., J. Universidad Católica de Salta(Bar-Ilan University and ABS Ventures. 2007.
[10] "Recuperación de Información de BD no estructuradas utilizando técnicas de minerías de textos que sean aplicables en el ámbito médico".
Aguilar J.P. Univ Nacional de Colombia, Bogotá. 2008
[11] "Clasification of text, automatic". In The Encyclopedia of Language and Linguistics. 2da Ed. Sebastianini, F. 2006.